UNIVERSITY
OF MICHIGAN
DEC 16 1959
PHYSICS
LIBRARY

Volume 63B

Number 2

Oct.–Dec. 1959

Journal of Research

of the

NATIONAL BUREAU OF STANDARDS

B. MATHEMATICS AND MATHEMATICAL PHYSICS

Journal of Research

of the

National Bureau of Standards

B. Mathematics and Mathematical Physics

OCT.–DEC. • 1959

VOLUME • 63B

NUMBER • 2

Editor: Chester H. Page

Associate Editors: F. L. Alt, E. L. Crow

Publication dates: Sept. 10, Dec. 10, 1959

JOURNAL OF RESEARCH

The National Bureau of Standards Journal of Research reports research and development in the fields of activity shown at right. Also included from time to time are survey articles and compilations of information on topics closely related to the Bureau's scientific and technical program.

The Journal is published in four separate sections as follows:

A. Physics and Chemistry

Contains papers of interest primarily to scientists working in these fields. Issued six times a year. Annual subscriptions: domestic, $4.00; foreign, $4.75.

B. Mathematics and Mathematical Physics

Presents studies and compilations designed mainly for the mathematician and the theoretical physicist. Issued quarterly. Annual subscription: domestic, $2.25; foreign, $2.75.

C. Engineering and Instrumentation

Reports research and development results of interest chiefly to the engineer and the applied scientist. Issued quarterly. Annual subscription: domestic, $2.25; foreign, $2.75.

D. Radio Propagation

Reports work in radio propagation physics and upper atmospheric research. Issued six times a year. Annual subscription: domestic, $4.00; foreign, $4.75.

U.S. DEPARTMENT OF COMMERCE
Frederick H. Mueller, Secretary

NATIONAL BUREAU OF STANDARDS
A. V. Astin, Director

Functions and Activities

The functions of the National Bureau of Standards are set forth in the Act of Congress, March 3, 1901, as amended. These include the development and maintenance of the national standards of measurement and the provision of means and methods for making measurements consistent with these standards; the determination of physical constants and properties of materials; the development of methods and instruments for testing materials, devices, and structures; advisory services to government agencies on scientific and technical problems; invention and development of devices to serve special needs of the Government; and the development of standard practices, codes, and specifications. The work includes basic and applied research, development, engineering, instrumentation, testing, evaluation, calibration services, and various consultation and information services. The Bureau also serves as the Federal technical research center in a number of specialized fields. The scope of activities of the National Bureau of Standards is suggested in the following listing of the divisions and sections engaged in technical work.

Washington, D.C.

Electricity and Electronics. Resistance and Reactance. Electron Devices. Electrical Instruments. Magnetic Measurements. Dielectrics. Engineering Electronics. Electronic Instrumentation. Electrochemistry.

Optics and Metrology. Photometry and Colorimetry. Optical Instruments. Photographic Technology. Length. Engineering Metrology.

Heat. Temperature Physics. Thermodynamics. Cryogenic Physics. Rheology. Molecular Kinetics. Free Radicals Research.

Atomic and Radiation Physics. Spectroscopy. Radiometry. Mass Spectrometry. Solid State Physics. Electron Physics. Atomic Physics. Neutron Physics. Radiation Theory. Radioactivity. X-ray. High Energy Radiation. Nucleonic Instrumentation. Radiological Equipment.

Chemistry. Organic Coatings. Surface Chemistry. Organic Chemistry. Analytical Chemistry. Inorganic Chemistry. Electrodeposition. Molecular Structure and Properties of Gases. Physical Chemistry. Thermochemistry. Spectrochemistry. Pure Substances.

Mechanics. Sound. Mechanical Instruments. Fluid Mechanics. Engineering Mechanics. Mass and Scale. Capacity, Density, and Fluid Meters. Combustion Controls.

Organic and Fibrous Materials. Rubber. Textiles. Paper. Leather. Testing and Specifications. Polymer Structure. Plastics. Dental Research.

Metallurgy. Thermal Metallurgy. Chemical Metallurgy. Mechanical Metallurgy. Corrosion. Metal Physics.

Mineral Products. Engineering Ceramics. Glass. Refractories. Enameled Metals. Constitution and Microstructure.

Building Technology. Structural Engineering. Fire Protection. Air Conditioning, Heating, and Refrigeration. Floor, Roof, and Wall Coverings. Codes and Safety Standards. Heat Transfer. Concreting Materials.

Applied Mathematics. Numerical Analysis. Computation. Statistical Engineering. Mathematical Physics.

Data Processing Systems. SEAC Engineering Group. Components and Techniques. Digital Circuitry. Digital Systems. Analog Systems. Applications Engineering.

• Office of Basic Instrumentation. • Office of Weights and Measures.

Boulder, Colorado

Cryogenic Engineering. Cryogenic Equipment. Cryogenic Processes. Properties of Materials. Gas Liquefaction.

Radio Propagation Physics. Upper Atmosphere Research. Ionosphere Research. Regular Prediction Services. Sun-Earth Relationships. VHF Research. Radio Warning Services. Airglow and Aurora. Radio Astronomy and Arctic Propagation.

Radio Propagation Engineering. Data Reduction Instrumentation. Radio Noise. Tropospheric Measurements. Tropospheric Analysis. Propagation-Terrain Effects. Radio-Meteorology. Lower Atmosphere Physics.

Radio Standards. High-Frequency Electrical Standards. Radio Broadcast Service. Radio and Microwave Materials. Electronic Calibration Center. Microwave Circuit Standards.

Radio Communication and Systems. Low Frequency and Very Low Frequency Research. High Frequency and Very High Frequency Research. Modulation Systems. Antenna Research. Navigation Systems. Systems Analysis. Field Operations.

Order all publications from the Superintendent of Documents U.S. Government Printing Office, Washington 25, D.C.

Use of funds for printing this publication approved by the Director of the Bureau of the Budget (June 24, 1958).

Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 62B, No. 2, October–December 1959

Applications of a Theorem on Partitioned Matrices

Emilie V. Haynsworth

(May 25, 1959)

A reduction formula for partitioned matrices is applied to block-stochastic matrices and other types of partitioned matrices in order to reduce the computation in finding the eigenvalues. Such matrices occur frequently in physics and chemistry, and the reduction formulas given here have been applied successfully in practical problems. In addition, some results of A. Brauer on stochastic matrices and of J. Williamson in partitioned matrices are generalized.

1. Introduction

In a previous paper,[1] a general reduction formula was given for certain partitioned matrices,

$$A=\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1t} \\ A_{21} & A_{22} & \cdots & A_{2t} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ A_{t1} & A_{t2} & \cdots & A_{tt} \end{pmatrix} \tag{1}$$

of order N, where the submatrices, A_{ij}, have dimensions $n_i \times n_j$, $i,j=1,2,\ldots t$, and

$$\sum_{i=1}^{t} n_i = N.$$

One of the special cases of this general theorem is of particular interest in practical applications to the problem of finding the eigenvalues of a matrix and, in fact, has been applied successfully to such problems. This special case is given below as theorem 1, since its proof is simpler than that for the general theorem and exhibits the transformation matrices needed for the reduction formula.

This result is used in section 3 to give generalizations of some theorems by A. Brauer [2] on stochastic matrices. Other applications are given in later sections.

As in the previous paper, for a given partitioning of a matrix A we shall call the submatrices, A_{ij}, the *blocks* of A and we shall write $A=(A_{ij})$. The elements of the blocks will be denoted by $a_{hk}^{(ij)}$, i.e.,

$$A_{ij}=(a_{hk}^{(ij)}).$$

Unless otherwise stated, the matrices will be arbitrary complex matrices. Also, since we will be dealing throughout with matrices $A=(A_{ij})$ of form (1), we will assume, unless otherwise indicated, that the statements and formulas given are true for $i,j=1,2,\ldots,t$.

2. Special Reduction Formula

For the sake of completeness we include a lemma from the previous paper [1] which is needed in the proofs of theorems 1 and 3. In this lemma we consider three cases where there are zeros in convenient places in the blocks of a partitioned matrix, A. In each of these cases A is reducible, and the proof consists merely in defining the permutation matrix which puts A into the reduced form,

$$\tilde{A}=\begin{pmatrix} C & O \\ D & T \end{pmatrix}, \tag{2}$$

where C and T are square matrices and O is a matrix composed entirely of zeros.

LEMMA: *Given a partitioned matrix A of order N with $n_1 \times n_j$ blocks A_{ij}:*

1. *If $n_i=n$ $(i=1,2,\ldots,t)$ and the blocks A_{ij} are lower triangular with elements $\lambda_{ij}^{(h)}$, $h=1,2,\ldots,n$, on the diagonal, A is similar to a matrix $\tilde{A}$, with blocks $\tilde{A}_{hh}=(\lambda_{ij}^{(h)})$, $h=1,2,\ldots,n$, on the diagonal, and zero blocks above the diagonal (this case is given by Williamson [3]).*

2. *If*

$$A_{ij}=\begin{pmatrix} C_{ij} & O \\ D_{ij} & T_{ij} \end{pmatrix}, \tag{3}$$

where all matrices T_{ij} are square, of order r, and the matrices C_{ij} are $(n_i-r)\times(n_j-r)$, then A has the form (2), where $C=(C_{ij})$ and $T=(T_{ij})$.

3. *If A_{ij} has the form (3) where the matrices T_{ij} are lower triangular, we will say A_{ij} is partially triangular. Then if A has blocks A_{ij} which are all partially triangular with submatrices, T_{ij} of order r, having elements $t_{ij}^{(h)}$, $h=1,2,\ldots,r$, on the diagonal, tr roots of A are roots of the r matrices, $(t_{ij}^{(h)})$.*

[1] E. Haynsworth, Reduction formulas for partitioned matrices (to be published).

[2] A. Brauer, Limits for the characteristic roots of a matrix, IV, Duke Math. J. 19, 75 (1952).

[3] J. Williamson, The latent roots of a matrix of special type, Bull. Am. Math. Soc. **37**, 585 (1931).

The proofs given below would follow in a corresponding manner if all blocks were transposed.

PROOF:

1. The rows and columns of A should be arranged in the order

$$\begin{aligned} &1, n+1, 2n+1, \ldots, (t-1)n+1; \\ &2, n+2, 2n+2, \ldots, (t-1)n+2; \\ &\qquad\qquad \cdots \\ &n, 2n, 3n, \qquad \ldots, \qquad tn. \end{aligned} \tag{4}$$

Then the new matrix $\tilde{A}$ will have the matrices $\tilde{A}_{hh}=(\lambda_{ij}^{(h)})$ on the diagonal and $\tilde{A}_{hk}=0$ for $i<j$, so its roots are the roots of the n matrices, $\tilde{A}_{hh}$.

2. Let $\sum_{i=1}^{k} n_i=N_k$. Then $N_1=n_1$, $N_t=N$. If we arrange the rows and columns of A in the following order:

$$\begin{aligned} &1, 2, \ldots, N_1-r, \\ &N_1+1, N_1+2, \ldots, N_2-r, \\ &\qquad\qquad \cdots \\ &N_{t-1}+1, N_{t-1}+2, \ldots, N_t-r, \\ &N_1-r+1, N_1-r+2, \ldots, N_1, \\ &\qquad\qquad \cdots \\ &N_t-r+1, N_t-r+2, \ldots, N_t, \end{aligned} \tag{5}$$

we have a new matrix A in which the matrices C_{ij} are together in the upper left corner, and the matrices T_{ij} are together in the lower right corner. So A will have the form (2) where $C=(C_{ij})$ and $T=(T_{ij})$.

3. Case 3 follows immediately now by first applying the permutation in 2 to the rows and columns of A and then applying the permutation in 1 to the rows and columns containing T.

THEOREM 1: *Suppose the blocks A_{ij} of the partitioned matrix given in (1) satisfy the equation*

$$A_{ij}X_j=X_iB_{ij}, \tag{6}$$

where B_{ij} is a square matrix of order r, $0<r\leqq n_i$, with strict inequality for at least one value of i, and X_i is an $n_i\times r$ matrix with a nonsingular matrix of order r, $X_i^{(1)}$, in the first r rows. Let the last n_i-r rows of X_i be $X_2^{(i)}$, and let

$$A_{ij}=\begin{pmatrix} A_{11}^{(ij)} & A_{12}^{(ij)} \\ A_{21}^{(ij)} & A_{22}^{(ij)} \end{pmatrix} \tag{7}$$

where $A_{11}^{(ij)}$ is square, of order r. Then A is similar to the matrix,

$$R=\begin{pmatrix} B & * \\ O & C \end{pmatrix} \tag{8}$$

where B is a partitioned matrix of order tr with blocks B_{ij}, as defined in (6), and C has blocks,

$$C_{ij}=(A_{22}^{(ij)}-X_2^{(i)}(X_1^{(i)})^{-1}A_{12}^{(ij)}) \tag{9}$$

with dimensions $(n_i-r)\times(n_j-r)$. (If either n_i or $n_j=r$, the corresponding block C_{ij} does not appear. By hypothesis not all n_i are equal to r, or else we would be left with the matrix $B=(B_{ij})$ which would be similar to A.)

Thus the roots of A are the roots of the smaller matrices B and C.

PROOF: Let P_i be a matrix of order n_i:

$$P_i=\begin{pmatrix} X_1^{(i)} & O \\ X_2^{(i)} & I_{n_i-r} \end{pmatrix}; \tag{10}$$

then

$$P_i^{-1}=\begin{pmatrix} (X_1^{(i)})^{-1} & O \\ -X_2^{(i)}(X_1^{(i)})^{-1} & I_{n_i-r} \end{pmatrix}, \tag{11}$$

where I_k represents an identity matrix of order k.

Since by (7) and (10)

$$A_{ij}P_j=\begin{pmatrix} [A_{11}^{(ij)}X_1^{(j)}+A_{12}^{(ij)}X_2^{(j)}] & A_{12}^{(ij)} \\ [A_{21}^{(ij)}X_1^{(j)}+A_{22}^{(ij)}X_2^{(j)}] & A_{22}^{ij} \end{pmatrix},$$

then by (6),

$$A_{ij}P_j=\begin{pmatrix} [X_1^{(i)}\cdot B_{ij}] & A_{12}^{(ij)} \\ [X_2^{(i)}\cdot B_{ij}] & A_{22}^{(ij)} \end{pmatrix},$$

and, using (11) and (9),

$$\tilde{A}_{ij}=P_i^{-1}A_{ij}P_j=\begin{pmatrix} B_{ij} & (X_1^{(i)})^{-1}A_{12}^{(ij)} \\ O & C_{ij} \end{pmatrix}. \tag{12}$$

So, if we let P be the direct sum of the matrices P_i,

$$P=\Sigma\cdot P_i=P_1\dot{+}P_2\dot{+}\ \ldots\ \dot{+}P_t, \tag{13}$$

and let

$$\tilde{A}=P^{-1}AP,$$

then $\tilde{A}$ has the blocks $\tilde{A}_{ij}$ given in (12), and the simultaneous permutation of rows and columns of A given by case 2 of the lemma above, will put it into form (8).

3. Generalizations of Brauer Theorems on Stochastic Matrices

A. Brauer [2] proved a number of interesting theorems about generalized stochastic matrices, i.e., matrices $A=(a_{ij})$ of order n such that

$$\sum_{j=1}^{n} a_{ij}=s, \qquad (i=1,2,\ldots n). \tag{14}$$

Such matrices have as one root, s, and the vector corresponding to this root is

$$U_n=(1,1,\ldots,1) \tag{14a}$$

The reduction formula of theorem 1 is applied in this section to matrices which can be partitioned into stochastic blocks, to give generalizations of Brauer's results.

If a matrix A of order N can be partitioned into rectangular $(n_i\times n_j)$ blocks, in each of which the row-sums are all equal, i.e.,

$$A_{ij}=(a_{hk}^{(ij)}), \qquad (h=1,2,\ldots,n_i, \quad k=1,2,\ldots,n_j)$$

with

$$\sum_{k=1}^{n_j} a_{hk}^{(ij)}=s_{ij}, \qquad (h=1,2,\ldots,n_i), \tag{15}$$

A is a *block-stochastic* matrix.

THEOREM 2: *If A is block-stochastic, (1) t roots of A are roots of the matrix $S=(s_{ij})$, and (2) the other $N-t$ are roots of the partitioned matrix of order $N-t$,*

$$C=(C_{ij}),$$

$$C_{ij}=(a_{hk}^{(ij)}-a_{1k}^{(ij)}), \qquad (h=2,3,\ldots,n_j, \quad k=2,3,\ldots,n_j). \tag{16}$$

If either n_i or n_j is 1, the block C_{ij} is omitted, so i and j do not necessarily take on all values $1, 2, \ldots, t$.

PROOF: Since, by (14a) and (15)

$$A_{ij}U_{n_j}=s_{ij}U_{n_i}, \tag{17}$$

statement 1 follows immediately, for by (17), if we set $r=1$ in (6) the matrices B_{ij} of theorem 1 are simply the elements s_{ij} given by (15). Then, if we subdivide A_{ij} as in (7), $A_{22}^{(ij)}$ consists of all the elements of the block A_{ij} except for those in the first row and column, i.e.,

$$A_{22}^{(ij)}=(a_{hk}^{(ij)}) \qquad (h=2,\ldots,n_i; \quad k=2,\ldots,n_j),$$

and

$$A_{12}^{(ij)}=(a_{1k}^{(ij)}) \qquad (k=2,\ldots,n_j).$$

Since by (17), the matrices X_i in (6) are the vectors U_{n_i}, we have

$$X_1^{(i)}=1, \qquad X_2^{(i)}=U_{n_i-1}.$$

Then, since $U_{n_i-1}A_{12}^{(ij)}$ is a matrix having in each row the elements

$$a_{12}^{(ij)},\ \ldots,\ a_{1,n_j}^{(ij)},$$

statement 2 follows from (9). For example, the following matrix arose from a problem in molecular physics: [4]

$$A=\left(\begin{array}{cc|cc|c|cccc} a & b & b & b & c & c & c & d & d \\ b & a & b & b & c & d & d & c & c \\ \hline b & b & a & b & d & c & d & c & d \\ b & b & b & a & d & d & c & d & c \\ \hline c & c & d & d & e & f & f & f & f \\ \hline c & d & c & d & f & e & f & f & g \\ c & d & d & c & f & f & e & g & f \\ d & c & c & d & f & f & g & e & f \\ d & c & d & c & f & g & f & f & e \end{array}\right).$$

The lines are drawn in to show how A is subdivided into stochastic blocks. Then by theorem 3, the roots of A are those of the 5×5 matrix,

$$C=\begin{pmatrix} a-b & 0 & c-d & c-d & d-c \\ 0 & a-b & c-d & d-c & c-d \\ c-d & c-d & e-g & 0 & 0 \\ c-d & 0 & f-g & e-f & g-f \\ 0 & c-d & f-g & g-f & e-f \end{pmatrix}$$

and those of the 4×4 matrix,

$$S=\begin{pmatrix} a+b & 2b & c & 2c+2d \\ 2b & a+b & d & 2c+2d \\ 2c & 2d & e & 4f \\ c+d & c+d & f & e+2f+g \end{pmatrix}.$$

It is easy to see that C can again be divided into stochastic blocks where $n_1=2$, $n_2=1$, $n_3=2$. So the five roots of C are those of the 2×2 matrix,

$$C_1=\begin{pmatrix} a-b & 2c-2d \\ c-d & e-g \end{pmatrix},$$

and the 3×3 matrix

$$C_2=\begin{pmatrix} a-b & c-d & 0 \\ 2c-2d & e-g & 0 \\ c-d & f-g & e-2f+g \end{pmatrix}.$$

Thus $e-2f+g$ is a root, and there are two degenerate roots since the matrix C_1 corresponds to the quadratic matrix which remains when the root $e-2f+g$ is factored out of C_2.

COROLLARY (THEOREM 2): *If A is a block-stochastic matrix with $n_i\times n_j$ blocks A_{ij}, and $P=\sum\cdot P_i$, $Q=\sum\cdot Q_i$ where P_i and Q_i are permutation matrices of*

[4] The author is grateful to J. Bradley for supplying this example.

order n_i, any matrix PAQ is also block-stochastic and has t roots in common with A.

PROOF: Since the row-sums in a given block would be unchanged by permutations within the block, this follows immediately from statement 1 of theorem 2.

Brauer (see footnote 2) proved the following theorems for certain special stochastic matrices:

1. Assume that

$$a_{ij}=c_j \qquad (j=2, 3, \ldots n; \quad i<j).$$

Then A has the same roots as the diagonal matrix

$$B=\text{diag}\,(a_{22}-c_2, a_{33}-c_3, \ldots, a_{nn}-c_n, s).$$

2. Assume

$a_{ij}=c$, $j\neq i$, and $a_{jj}=b$

$$(j=j_1, j_2, \ldots, j_r).$$

Then $(b-c)$ is a root of multiplicity r.

We generalize these results in the following theorem for block-stochastic matrices:

THEOREM 3: *Let A be a block-stochastic matrix with submatrices $A_{ij}=(a_{hk}^{ij})$ having constants c_k above the diagonal in the last r columns of each block, $0<r<\min n_i$, i.e.,*

$$a_{hk}^{(ij)}=c_k^{(ij)} \qquad (k>n_{ij}-r; \quad h<k+n_i-n_j). \quad (18)$$

Then tr of the roots of A are roots of the matrices

$$Q_{kk}=(a_{kk}^{(ij)}-c_k^{(ij)}) \quad (k=n_j-r+1, \ldots, n_j). \quad (19)$$

(The second of Brauer's theorems comes as a special case of the above if we permute the rows and columns so that the columns $j_1, j_2, \ldots, j_r$ are the last r columns of the matrix. Also it isn't necessary that all the nondiagonal elements of a column be equal as in Brauer's hypothesis.)

PROOF: By theorem 2, N-t roots of A are roots of $C=(C_{ij})$. Here we have $i, j=1,2, \ldots, t$, since by hypothesis, $0<r<\min_i n_i$, and thus $\min_i n_i>1$.

But by (16) and (18), C_{ij} is a partially triangular matrix with elements $a_{kk}^{(ij)}-c_k^{(ij)}$ on the diagonal of the triangular portion, i.e., for $k=n_j-r+1, \ldots, n_j$. Then, using case 3 of the lemma above, we can permute the rows and columns of $\bar{A}$ so that $\bar{A}$ is similar to the matrix

$$R=\begin{pmatrix} P & * \\ O & Q \end{pmatrix}$$

when $Q=(Q_{hk})$ is a matrix of order tr having square blocks of order t, defined by (15), on the diagonal, and $Q_{hk}=0$ for $k<h$.

4. Other Applications

In this section we will mention several other types of partitioned matrices, $A=(A_{ij})$, such that

$$PAP^{-1}=\begin{pmatrix} B & * \\ O & C \end{pmatrix}, \qquad (20)$$

where P is the direct sum of transformation matrices P_i, and $B=(B_{ij})$, $C=(C_{ij})$, are partitioned matrices which will be defined in each case. We also give a case where the transformation PAP^{-1} produces a real matrix from a complex one.

Since the proof in each case is the same and consists merely in performing the indicated matrix multiplications, it will not be given in detail.

4.1. Block-Circulant Matrices

This is the case where each block A_{ij} is a circulant matrix of order n. The result (21) is given by Williamson.[3]

The roots of A are roots of the n matrices of order t,

$$(\lambda_{ij}^h)=\left(\sum_{k=1}^{n} a_{hk}^{(ij)}\epsilon_h^k\right) \qquad (h=1, 2, \ldots, n) \qquad (21)$$

where ϵ_h is one of the nth roots of unity.

4.2. Block-Singular Matrices

If A is singular, of rank $n-1$, there exists a vector $x=(x_1, x_2, \ldots, x_n)$, with components not all zero, such that

$$AX=X\cdot O.$$

We may assume, without loss of generality, that $x_1=1$. Then, as in theorem 1, if $X=1+X_2$, we can let

$$P=\begin{pmatrix} 1 & O \\ -X_2 & I_{n-1} \end{pmatrix} \qquad (22)$$

and, applying the transformation (16), we can show that the nonzero roots of A are roots of

$$C=(a_{ij}-x_i a_{ij}) \qquad (i,j=2, \ldots, n).$$

If A has rank less than $n-1$, the same type of transformation can be applied to the matrix C and this can be continued until a nonsingular matrix is obtained. Or, we may be able to partition A into blocks so that

$$A_{ij}X_j=X_i\cdot O.$$

Then if we let $P=\sum P_i$, where each P_i has the form (22), and apply transformation (20), we show

that the roots of A are the roots of

$$B=(B_{ij})=(O) \quad \text{of order } t$$

and

$$C=(C_{ij})$$

where

$$C_{ij}=(a_{hk}^{(ij)}-x_h^{(i)}\,a_{1k}^{(ij)})$$
$$(h=2,\ldots,n_i; \qquad k=2,\ldots,n_j).$$

4.3. Block Tridiagonal Matrices

Let

$$T(a)=\begin{pmatrix} a & 1 & 0\ldots 0 & 0 \\ 1 & a & 1\ldots 0 & 0 \\ 0 & 1 & a\ldots 0 & 0 \\ \cdot & \cdot & \cdot\quad\cdot & \cdot \\ \cdot & \cdot & \cdot\quad\cdot & \cdot \\ \cdot & \cdot & \cdot\quad\cdot & \cdot \\ 0 & 0 & 0\ldots a & 1 \\ 0 & 0 & 0\ldots 1 & a \end{pmatrix} \tag{23}$$

It is well known[5] that such matrices have characteristic roots which may be written in the form

$$\lambda_k=a+2\cos(k\pi/n+1) \qquad (k=1,2,\ldots,n) \tag{24}$$

and, if we let $\theta_k=k\pi/n+1$ it is easy to show that the characteristic vector corresponding to λ_k has components

$$(\sin\theta_k,\ \sin 2\theta_k,\ \ldots,\ \sin n\theta_k) \tag{25}$$

since the trigonometric identities

$$a\sin\theta_k+\sin 2\theta_k=(a+2\cos\theta_k)\sin\theta_k,$$

$$\sin(m-1)\theta_k+a\sin m\theta_k+\sin(m+1)\,\theta_k$$
$$=(a+2\cos\theta_k)\sin m\theta_k \quad (m=2,\ldots,n-1),$$

$$\sin(n-1)\,\theta_k+a\sin n\theta_k=(a+2\cos\theta_k)\sin n\theta_k$$

hold for each θ_k as defined above.

So if $A=(A_{ij})$ where the blocks A_{ij} are all tridiagonal,

$$A_{ij}=T(a_{ij}),$$

these blocks can be simultaneously reduced to triangular form using the matrix of characteristic vectors (25), since these are independent of the elements a_{ij}. Thus, according to a theorem by Williamson,[4] the roots of A are those of the n matrices,

$$(a_{ij}+2\cos(\pi/n+1)),\ \ldots,$$
$$(a_{ij}+2\cos(n\pi/n+1)). \tag{26}$$

[4] D. E. Rutherford, Some continuant determinants arising in physics and chemistry, II, Proc. Roy. Soc. Edinburgh **63**, III (1952).

If, in addition, the matrix (a_{ij}) is stochastic, with row sum s, by theorem 1, A has as n of its roots the number $s+(1/n+1)2t\cos k\pi$ $(k=1,\ldots,n)$. The other $n(t-1)$ roots are degenerate with multiplicity n, since the matrices C_{ij} in (5) become, by (22) for any value of n,

$$(a_{ij}-a_{1j}) \qquad (i,j=2,\ \ldots\ t). \tag{27}$$

4.4. Complex Matrices Similar to Real Matrices

$A=(A_{ij})$ where each block is a 2×2 matrix defined as follows:

$$A_{ii}=\begin{pmatrix} a_i+\sqrt{-1}\,b_i & 0 \\ 0 & a_i-\sqrt{-1}\,b_i \end{pmatrix} \qquad (i=1,\ldots,k).$$

(We use $\sqrt{-1}$ instead of i to avoid confusion with the subscript.)

$$A_{ij}=\begin{pmatrix} a_{ij} & b_{ij} \\ b_{ij} & a_{ij} \end{pmatrix} \qquad (i,j=1,2,\ldots,k;\quad i\pm j)$$

$$A_{ij}=\begin{pmatrix} a_{ij} & b_{ij} \\ a_{ij} & b_{ij} \end{pmatrix} \qquad (i=1,2,\ldots,k;\quad j=k+1,\ldots,t)$$

$$A_{ij}=\begin{pmatrix} a_{ij} & a_{ij} \\ b_{ij} & b_{ij} \end{pmatrix} \qquad (i=k+1,\ldots,t;\quad j=1,\ldots,k)$$

and A_{ij} an arbitrary real matrix $(i,j=k+1,\ldots,t)$. Then,

$$P_i=\begin{pmatrix} \dfrac{1+\sqrt{-1}}{2} & \dfrac{1-\sqrt{-1}}{2} \\ \dfrac{1-\sqrt{-1}}{2} & \dfrac{1+\sqrt{-1}}{2} \end{pmatrix} \qquad (i=1,2,\ldots,k),$$

$$P_i=I_2 \qquad (i=k+1,\ldots,n).$$

$$P_iA_{ii}P_i^{-1}=\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \qquad (i=1,\ldots,k),$$

$$P_iA_{ij}P_j^{-1}=A_{ij}, \qquad (i,j=1,\ldots,k;),$$

$$P_iA_{ij}=A_{ij} \qquad (i=1,\ldots,k,\quad j=k=1,\ldots,t),$$

$$A_{ij}P_j^{-1}=A_{ij} \qquad (i=k+1,\ldots,t;\quad j=1,\ldots k),$$

Since

$$\widetilde{A}=PAP^{-1} \tag{28}$$

has the same elements as A except in the 2×2 blocks on the diagonal for $i=1, 2, \ldots, k$, and the elements in these blocks are real. (Matrices of this type also arose from an applied problem at the National Bureau of Standards.)

5. Applications to Determinants and Systems of Equations

The transformations given here for a partitioned matrix $A=(A_{ij})$ can obviously be applied to finding the determinant of such a matrix since if

$$\tilde{A}=PAP^{-1}=\begin{pmatrix} B & D \\ O & C \end{pmatrix}, \tag{29}$$

$$|\tilde{A}|=|A|=|B|\cdot|C|.$$

These transformations could also be profitably applied to the solution of a system of linear equations with coefficient matrix A; i.e., the system of equations

$$AX=K \tag{30}$$

is, by (29), the same as

$$(P^{-1}\tilde{A}P)X=K$$

or

$$\tilde{A}\tilde{X}=\tilde{K}$$

where

$$PK=\tilde{K}$$

and

$$PX=\tilde{X}. \tag{31}$$

Then if $X=X_1+X_2$ and $K=K_1+K_2$ are partitioned to correspond with the dimensions of B and C, we can first solve the system

$$C\tilde{X}_2=\tilde{K}_2 \tag{32}$$

and then the system

$$B\tilde{X}_1=\tilde{K}_1-D\tilde{X}_2. \tag{33}$$

So, solving these two systems of linear equations and the comparatively simple system (31) (which can be partitioned as P is partitioned), we have the solution of (29).

Such a transformation would be especially valuable in solving a large system on an electronic computer, where the length of time for solution goes up on the order of n^3. Thus, if we cut the dimensions in half by transformation (29), systems (32) and (33) would each take one-eighth as long to solve as the original system, and if we assumed only that P could be partitioned into two parts the total solution time would be at most one-half that of the original system. In many cases the time could be cut to approximately one-fourth or less.

Washington, D.C. (Paper 63B2–9)

Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 63B, No. 2, October–December 1959

Capacity Requirement of a Mail Sorting Device: II[1]

A. J. Goldman

(July 14, 1959)

The combinatorial analysis of a mathematical model of a sorting device suggested by S. Henig is completed. The relevant parameters are r, the number of destinations for the mail, and k, the number of letters entering the device during each cycle of operation. The capacity of the device, if it is never to jam, should be between rk and $rk-(r-1)$ inclusive; arguments indicating that the latter value is preferable are given.

1. Introduction

We deal with a highly idealized mathematical model of a mail sorting device suggested by S. Henig (NBS Electronic Instrumentation Section). For our purposes the operation of the device can be described[2] as follows. Mail (to any of r destinations) enters the system. After k letters have entered, the device "asks itself" to which destination it contains the *most* letters.[3] All letters to this *predominant destination* are then *dropped out* of the device, another k letters enter, and the process continues. The device is said to *jam* if, after a dropout, it is still so "full" that entrance of the next k letters would cause an "overflow."

Investigation of the most appropriate capacity for such a device leads to mathematically interesting problems of two general types. If the capacity is to be so chosen that jamming *never* occurs, then the problems are essentially combinatorial in nature and can be treated mainly by "counting" methods. If, however, the object is the more modest one of keeping the frequency of jamming down to some specified "tolerable level," then rather difficult probabilistic problems arise; this is due to the fact[4] that the behavior of the device constitutes a Markov chain with a great many states, governed by the probabilistic distribution of mail by destinations.

The present paper completes an earlier analysis[5] of the *combinatorial* questions, which showed that

$$(r-1)(k-1)+k=rk-(r-1) \tag{1}$$

is a possible capacity for the device, and (if the device begins operation empty) is in fact the *minimum* capacity. This left open the possibility of gaining some advantage by using a *greater* capacity than that given in (1) and beginning operation with the device partly or entirely filled. It will be proved below that no capacity in excess of rk should be employed, and that any advantage arising from a capacity greater than $rk-(r-1)$ would be only transient.

The author acknowledges several improvements in the exposition suggested by J. R. Rosenblatt. Unfortunately a somewhat more formal treatment than that of the previous paper seemed required by the greater complexity of the analysis below.

2. Statement of Results

In order to describe our results, it is convenient to define

$$x(t)=\text{number of letters in the device just before the } t\text{th dropout}, \tag{2}$$

and also to give a precise definition of "capacity."

A non-negative integer C will be called a *capacity* for the device if, should the device begin operation containing no more than C letters, it can never subsequently contain more than C letters under the dropout rule described above; symbolically,

$$x(t_0)\leq C \quad \text{implies} \quad x(t)\leq C \qquad \text{for all} \quad t\geq t_0. \tag{3}$$

This condition is clearly necessary to insure that the device never jams.

THEOREM 1. *C is a capacity if and only if*

$$rk-(r-1)\leq C.$$

A capacity C will be called *efficient* if there is at least one set of initial contents for the device, with no more than C letters, such that the device might *possibly* contain C letters again at some later time; symbolically, for *some* set of initial contents,

$$x(t_0)\leq C \quad \text{and} \quad x(t)_{\max \text{ poss}}=C \quad \text{for} \quad t>t_0. \tag{4}$$

If this condition is not satisfied, then the full capacity C of the device will not actually be needed after the start of operation, a wasteful situation.

THEOREM 2. *C is an efficient capacity if and only if*

$$rk-(r-1)\leq C\leq rk.$$

A capacity C will be called *essentially efficient*[6] if there is at least one set of initial contents with no more than C letters such that the device with non-zero probability will contain C letters infinitely

---

[1] Part of a project sponsored by the Post Office Department, Office of Research and Development.

[2] The physical device can also operate under dropout rules other than the one described below.

[3] A rule for breaking "ties" between destinations is also required.

[4] Pointed out (unpublished memorandum, July 1956) by J. R. Rosenblatt (NBS Statistical Engineering Laboratory).

[5] B. K. Bender and A. J. Goldman, Capacity requirement of a mail sorting device, J. Research NBS **62**, 171 (1959) RP2948.

[6] Use of the term "essential" here, and of the subscript "ess" in (5), was suggested by a roughly similar usage in measure theory.

often [7]; symbolically, for some set of initial contents,

$$x(t_0) \leq C \quad \text{and} \quad x(t)_{\max \text{ ess}} = C \quad \text{for} \quad t > t_0. \tag{5}$$

This definition is strengthened to that of a *uniformly essentially efficient capacity* by requiring that the stated condition hold for *all* initial contents of no more than C letters; i.e., that

$$x(t_0) \leq C \quad \text{implies} \quad x(t)_{\max \text{ ess}} = C \quad \text{for} \quad t > t_0. \tag{6}$$

THEOREM 3. *The only essentially efficient capacity is* $rk-(r-1)=C$; *it is also uniformly essentially efficient.*

The proofs of these theorems are given in section 3. Section 4 contains a very crude probabilistic estimate related to theorem 3.

3. Proofs

3.0. Preliminaries

It is helpful to define

$$M(t) = \max_i x_i(t)$$

where

$x_i(t)$ = number of letters to the ith destination in the device just before the tth dropout.

LEMMA 1. $M(t) > k$ *implies* $x(t+1) < x(t)$.
LEMMA 2. $x(t) > rk$ *implies* $x(t+1) < x(t)$.
LEMMA 3. $M(t) = k$ *implies* $x(t+1) = x(t)$.
LEMMA 4. $M(t) \geq k$ *implies* $x(t+1) \leq x(t)$.
LEMMA 5. $x(t) \geq rk-(r-1)$ *implies* $x(t+1) \leq x(t)$.
LEMMA 6. $x(t_0) \leq rk-(r-1)$ *implies* $x(t) \leq rk-(r-1)$ *for all* $t \geq t_0$.
LEMMA 7. $x(t_0) \geq rk-(r-1)$ *implies* $x(t) \leq x(t_0)$ *for all* $t \geq t_0$.

Lemma 1 is proved by observing that if $M(t) > k$, then more than k letters leave the device in the tth dropout, whereas only k letters enter before the $(t+1)$-st dropout. Since there are only r destinations, $x(t) > rk$ implies $M(t) > k$, and so lemma 2 follows from lemma 1. Lemma 3 is an obvious consequence of the dropout rule, and lemma 4 is obtained by combining lemmas 1 and 3.

To prove lemma 5, note that since there are only r destinations,

$$x(t) \geq rk-(r-1) > r(k-1) \quad \text{implies} \quad M(t) > k-1,$$

so that lemma 4 can be applied. Lemma 6 merely states the fact [5] that $rk-(r-1)$ is a capacity. In proving lemma 7, two possibilities must be considered. If, on the one hand,

$$x(t) \geq rk-(r-1) \qquad \text{for all} \quad t \geq t_0,$$

then by lemma 5, $x(t) \leq x(t_0)$ for all $t \geq t_0$. If, on the other hand, there is a $t_1 > t_0$ such that

$$x(t) \geq rk-(r-1) \qquad \text{for} \quad t_0 \leq t < t_1$$

but

$$x(t_1) < rk-(r-1),$$

then by lemma 5, $x(t) \leq x(t_0)$ for $t_0 \leq t < t_1$, and by lemma 6

$$x(t) \leq rk-(r-1) \leq x(t_0) \qquad \text{for all} \quad t \geq t_1,$$

so that again

$$x(t) \leq x(t_0) \qquad \text{for all} \quad t \geq t_0.$$

3.1. Proof of Theorem 1

To prove the "only if" statement, suppose $C < rk-(r-1)$ and $x(t_0) \leq C$. If, on the one hand, $x(t_0) = 0$, then (see *op. cit.* footnote 5, p. 79) there is a sequence of possible events leading to a situation in which

$$x(t) = rk-(r-1) > C; \qquad t \leq t_0 + (r-1)(k-1) + 2.$$

This violates condition (3) and so C is not a capacity. If, on the other hand, $x(t_0) > 0$, then a sequence of possible events can be constructed leading to a situation in which

$$x(t_1) = 0; \qquad t_1 \leq t_0 + C. \tag{7}$$

To do this, choose the k letters entering the device between the tth and $(t+1)$-st dropouts, for $t_0 \leq t < t_1$, to consist entirely of letters to some one destination for which $x_i(t) > 0$. Then $M(t) > k$ for $t_0 \leq t < t_1$, and so by lemma 1 the number of letters in the device steadily decreases until condition (7) is satisfied. From this point we may argue as in the case $x(t_0) = 0$ (now using t_1, instead of t_0), again obtaining the conclusion that C is not a capacity.

To prove the "if" statement, suppose

$$C \geq rk-(r-1) \quad \text{and} \quad x(t_0) \leq C.$$

If $x(t_0) \leq rk-(r-1)$, then by lemma 6

$$x(t) \leq rk-(r-1) \leq C \qquad \text{for all} \quad t \geq t_0,$$

while if $x(t_0) \geq rk-(r-1)$, then by lemma 7

$$x(t) \leq x(t_0) \leq C \qquad \text{for all} \quad t \geq t_0;$$

in either case condition (3) is satisfied, and so C is a capacity.

[7] If this condition is not satisfied, then with probability one the device will, from some point on, fail to use its full capacity.

3.2. Proof of Theorem 2

To prove the "if" statement, suppose

$$rk-(r-1)\leq C\leq rk.$$

The initial contents of the device can clearly be chosen so that $x(t_0)=C$ and $M(t_0)=k$. Among the destinations with $x_i(t_0)=k$, select the one "preferred" by the tie-breaking rule at the t_0-th dropout; let k letters to this destination enter the device between the t_0-th and (t_0+1)-st dropouts. Then $M(t_0+1)=k$; among the destinations with $x_i(t_0+1)=k$, select the one preferred by the tie-breaking rule at the (t_0+1)-st dropout, and let k letters to *this* destination enter the device between the (t_0+1)-st and (t_0+2)-nd dropouts, etc. This construction results in $M(t)=k$ for all $t\geq t_0$, so that, by lemma 3,

$$x(t)=x(t_0)=C \qquad \text{for all} \quad t\geq t_0.$$

Hence C is an efficient capacity.

The "only if" statement will be proved next. By theorem 1, no $C<rk-(r-1)$ can be a capacity; thus no such C can be an efficient capacity. Suppose $C>rk$ and $x(t_0)\leq C$. There are three possibilities to be considered. If $x(t_0)\leq rk-(r-1)$, then by lemma 6

$$x(t)\leq rk-(r-1)\leq rk<C \qquad \text{for all} \quad t\geq t_0$$

so that C is not an efficient capacity. If

$$rk-(r-1)<x(t_0)\leq rk,$$

then by lemma 7

$$x(t)\leq x(t_0)\leq rk<C \qquad \text{for all} \quad t\geq t_0,$$

so that C is not an efficient capacity. Finally, if $x(t_0)>rk$, then by lemma 2 there is a $t_1>t_0$ (with $t_1\leq t_0+C-rk$) such that

$$rk<x(t)<x(t_0)\leq C \qquad \text{for} \quad t_0<t<t_1$$

and

$$x(t_1)\leq rk;$$

the arguments for the first two possibilities can be applied to t_1 to yield

$$x(t)\leq x(t_1)\leq rk<C \qquad \text{for all} \quad t\geq t_1$$

so that $x(t)<C$ for all $t>t_0$ and again C is not an efficient capacity.

3.3. Proof of Theorem 3

First, $rk-(r-1)$ *is* a uniformly essentially efficient capacity. To see this, note that the construction used above to prove the "only if" part of theorem 1 shows that if

$$x(t_0)\leq rk-(r-1)$$

then with probability one [8] there is a $t_1>t_0$ for which

$$x(t_1)=rk-(r-1);$$

by the very same argument, with probability one there is a $t_2>t_1$ for which

$$x(t_2)=rk-(r-1),$$

and so on. In fact, for $rk-(r-1)$ the phrase "with nonzero probability" in the definition of "uniformly essentially efficient capacity" could actually have been strengthened to "with probability one."

Second, no $C>rk-(r-1)$ is an essentially efficient capacity. To prove this, observe that by theorem 2 attention can be confined to the case

$$rk-(r-1)<C\leq rk.$$

Suppose $x(t_0)\leq C$. If

$$x(t_1)\leq rk-(r-1) \qquad \text{for some} \quad t_1\geq t_0,$$

then by lemma 6

$$x(t)\leq rk-(r-1)<C \qquad \text{for all} \quad t\geq t_1,$$

so that the device cannot contain C letters infinitely often. Thus we may assume that

$$x(t)>rk-(r-1) \qquad \text{for all} \quad t\geq t_0, \tag{8}$$

so that, by lemma 5,

$$x(t+1)\leq x(t) \qquad \text{for all} \quad t\geq t_0. \tag{9}$$

If $x(t_0)<C$, then by (9)

$$x(t)\leq x(t_0)<C \qquad \text{for all} \quad t\geq t_0,$$

so that the device can never contain C letters. Therefore assume that $x(t)_0=C$. If

$$M(t_1)>k \qquad \text{for some} \quad t_1\geq t_0,$$

then by lemma 1 and (9)

$$x(t)\leq x(t_1+1)<x(t_1)\leq x(t_0)=C \qquad \text{for all} \quad t>t_1,$$

so that the device cannot contain C letters infinitely often. Therefore assume

$$M(t)\leq k \qquad \text{for all} \quad t\geq t_0.$$

From (8) and the fact that there are only r destinations it follows that $M(t)\geq k$, and so $M(t)=k$. Thus the only possibility for the device to contain C letters infinitely often is given by

$$x(t)=C, \qquad M(t)=k \qquad \text{for all} \quad t\geq t_0;$$

this clearly has probability zero.

[8] The construction shows that if $x(t_0)\leq rk-(r-1)$, then for any block of $[(r-1)(k-1)+2+rk-(r-1)]$ values of $t\geq t_0$, there is a nonzero probability (independent of the particular block) that $x(t)=rk-(r-1)$ at least once.

4. A Probabilistic Estimate

Suppose C is an efficient capacity which is *not* essentially efficient; i.e. (according to theorems 2 and 3),

$$rk-(r-1)<C\leq rk.$$

We know that with probability one, there is a $t_1\geq t_0$ such that

$$x(t)<C \qquad \text{for all} \quad t>t_1.$$

It is of interest to estimate how early the first such t_1 (the last moment at which the full capacity of the device is used) might occur. In this primarily combinatorial paper, we will be content with an exceedingly crude probabilistic estimate.

Denote by E the event

$$x(t_0)=C \qquad \text{and} \quad M(t_0)=k.$$

By the proof of theorem 3, $t_1=t_0$ unless E occurs. Therefore the function of t_1 to be estimated is

$$P(t_1)=\text{Prob}\{x(t)=0 \quad \text{for some} \quad t>t_1|E\}, \quad (10)$$

where Prob$\{A|B\}$ denotes the conditional probability of A given B.

Let the relative frequencies (or probabilities) of letters to the different destinations be denoted by

$$0<f_1\leq f_2\leq \ \ldots \ \leq f_r,$$

let

$$C=r(k-1)+s+1; \qquad 0<s<r,$$

and let

$$F_s=f_1+f_2+ \ \ldots \ +f_s.$$

It will first be shown that, for any $t>t_0$,

$$\text{Prob}\{x(t)=C, \quad M(t)=k|x(t-1)=C, \quad M(t-1)=k\} \leq (1-F_s)^k \quad (11)$$

Suppose

$$x(t-1)=C, \qquad M(t-1)=k.$$

Since

$$x(t-1)=C>r(k-1)+s=(r-s)(k-1)+sk,$$

it follows that $x_i(t-1)=k$ for at least $s+1$ destinations. The letters to some one of these destinations leave the device in the $(t-1)$-th dropout, and $M(t)>k$ if any of the k letters which next enter the device go to any of the remaining s destinations with k letters each. Since the sum of any s frequencies f_i is at least F_s, the inequality (11) holds.

By the proof of theorem 3, for any one $t>t_0$, both $x(t-1)=C$ and $M(t-1)=k$ hold if and only if

$$x(t')=C \qquad \text{and} \qquad M(t')=k \qquad \text{for} \quad t_0\leq t'<t.$$

Thus the probability on the left side of (11) can be rewritten

$$\text{Prob}\{x(t)=C, \ M(t)=k|x(t')=C, \ M(t')=k \text{ for } t_0\leq t'<t\},$$

so that the inequality (11) yields

$$\Pi \ \text{Prob}\{x(t)=C, \ M(t)=k|x(t')=C, \ M(t')=k t_0<t\leq t_1 \text{ for } t_0\leq t'<t\} \leq (1-F_s)^{k(t_1-t_0)}.$$

This last product is equal to

$$\text{Prob}\{x(t)=C, \quad M(t)=k \quad \text{for} \quad t_0<t\leq t_1|E\},$$

which, according to the proof of theorem 3, is an upper bound for $P(t_1)$ and so we have proved that

$$P(t_1)\leq (1-F_s)^{k(t_1-t_0)}. \quad (12)$$

If for example $k=2$ and $F_s=0.001$, then after 500 dropouts the probability that the device will make any further use of its full capacity is less than 0.50.

WASHINGTON, D.C. (Paper 63B2-10)

Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 63B, No. 2, October–December 1959

Analytic Comparison of Suggested Configurations for Automatic Mail Sorting Equipment[1]

B. K. Bender and A. J. Goldman

(June 5, 1959)

Analytic methods are developed to aid in determining the equipment configuration which achieves sorting of outgoing mail at a given (required) rate at least cost. The techniques are applied to a specific numerical problem; several of the suggested configurations are quickly eliminated, and a "hybrid" of two of the proposed configurations is found which comes within four percent of optimum (if a certain pair of parameters is chosen correctly).

1. Introduction

A number of plausible configurations for automatic mail sorting equipment have been suggested. In this report methods for the comparative study of such configurations are developed, with the object of determining the configuration which achieves sorting of mail at a given (required) rate at least cost.

The solution to this "optimization" problem of course depends upon the distribution (by destinations) and volume of mail at the sorting installation, and also on the operating parameters and costs associated with the various components of the sorting system. Therefore, in order to present our methods in the most comprehensible way (i.e., in the context of a specific numerical application), it was necessary to work with a definite set of realistic "input" data. These inputs are described in sections 3 to 6 and appendix A. We wish to emphasize that we work with these particular data only to illustrate the methods employed; the reader who follows out the calculations for this specific case should have little trouble carrying out the corresponding calculations for other data.

The results of the analyses are summarized in section 2. In brief, we are able (*i*) to rule out several of the proposed configurations, (*ii*) to show that one of the proposed configurations will (if a certain parameter is choosen correctly) come within 8 per cent of optimum, and (*iii*) to exhibit a "hybrid" of two of the proposed configurations which (if a certain pair of parameters is chosen correctly) will come within 4 percent of optimum.

We also determine how accurately the parameters in (*ii*) and (*iii*) must be chosen. In view of (*iii*) analysis of more complicated configurations did not seem worthwhile.

For a comparative study, it is necessary only to consider costs which vary appreciably from one configuration to another. Our analysis is set up so that these variable costs are all absorbed into costs associated with (*a*) the devices which inject mail into the sorting system (we will generally use the term *loading complex* for such a device, instead of the longer "distributor loader complex"), (*b*) the receptacles (*bins*) in which the mail ends after passing through each stage of the system, and (*c*) the operation (*sweeping*) of removing the mail from these bins.

A natural method of determining the cost of any proposed configuration, therefore, is to determine the numbers of loading complexes, bins, and sweepers involved, and then to multiply each of these numbers by the appropriate cost coefficient and add up the resulting partial costs. We might call this the *add up* method; it is described in section 7 and at the end of section 8. We have also developed a *follow through* method based on the idea of following an individual letter through the sorting system, charging it an appropriate fraction of the cost of each loading complex through which it passes, each loading complex by which it passes (and thus prevents from operating at maximum rate), each bin it occupies, and each sweep it receives. The follow through method is not strictly applicable to some configurations (though it does seem to provide initial approximations which may shorten some of the work of the "add up" method); its main advantage is that it permits a much closer estimation[2] of the *theoretical minimum cost* described in the next paragraph.

An important step in the analysis is the derivation of a theoretical minimum cost (more precisely, a "lower bound") which is independent of the configuration. This cost is *minimum* in the sense that "any actual configuration must cost at least this much" and is *theoretical* in the sense that "no actual configuration can cost quite this little"; it therefore provides a *yardstick* against which the costs of specific configurations can be measured, the deviation from optimality of a configuration can be assessed, and the permissibility of plausible approximations can be checked. A rough estimate of such a minimum cost is given in section 8 (without using the specific mail distribution); in section 9 the "follow through" method is employed to derive a significantly more accurate result.[2]

[1] The preparation of this paper was sponsored by the Post Office Department, Office of Research and Engineering.

[2] The importance of this is discussed in section 8.

The detailed analyses of the configurations studied are given in sections 10 to 14 and appendixes B and C. We hope that our rather full exposition of this material will prove helpful to the reader faced with the problem of carrying out similar analyses.

2. Summary of Results

From our *specific numerical results* (which are valid only for the particular data and assumptions used in our analysis) we can infer certain *qualitative results*, believed to be valid for post-offices (*i*) for which the sorting system must be able to handle about 1,000,000 letters/hr or more and (*ii*) in which local mail is removed before the outgoing sort.

Of these two types of results, the qualitative ones are more informative, and so we list them first:

(*a*) The simplex scheme and Christmas Tree scheme [3] can be eliminated from further consideration.

(*b*) The residue scheme (if properly chosen) is quite good; some multiple input schemes may possibly be competitive.

(*c*) A promising approach is to combine the basic ideas of the residue and multiple input schemes; the resulting "hybrid" scheme has lower cost than either of the two original schemes if the relevant parameters are correctly chosen.

We turn now to the specific numerical results. The "inputs" leading to these results are given in detail in sections 3–6 and appendix A; at this point we only add to (*i*) and (*ii*) above the facts that (*iii*) only costs which vary between different equipment configurations were considered, (*iv*) the use of automatic transfer equipment was not considered, and (*v*) the possibility of memory-sharing between oading complexes was not considered.

Condition (*iii*) is a natural one, since we are primarily interested in comparing different configurations. Auxiliary calculations (not given in this paper) show that the qualitative results reported above do not depend on (*iv*) and (*v*).

As noted in section 1, a "theoretical minimum cost" is used as a yardstick against which the desirability of any proposed configuration can be measured. For this purpose we use the ratio

$$R=\frac{(\text{Cost of proposed config.})-(\text{Theoretical minimum})}{(\text{Theoretical minimum})} \quad (2.1)$$

"Near-optimal" configurations are those with *small* values of R. In the notation of the body of the report, (2.1) becomes

$$R=(C-C^*_{\min})/C^*_{\min}. \quad (2.2)$$

Our specific numerical results are as follows:

(*a*) The theoretical minimum (yearly variable) cost is about $2,050,000.

(*b*) For the simplex scheme, R is *about 160 percent;* i.e., the scheme costs about 2.6 times the theoretical minimum cost.

(*c*) For the Christmas Tree scheme, R is *at least 69 percent;* i.e., the scheme costs at least 1.69 times the theoretical minimum cost.

(*d*) For the best residue scheme, R is *about 8 percent.*

(*e*) There are a great many ways in which the basic ideas of the residue and multiple input schemes can be combined. An analysis of the full range of possibilities would be beyond the scope of our study. We have, however, examined a relatively simple class of systems of this type, in which each subsystem involves only two loading complexes in series; the optimal configuration within this class has R *about 4 percent*, which (*i*) compares favorably with all other configurations studied, and (*ii*) indicates that investigation of more complicated systems would not be worthwhile.

(*f*) All multiple input schemes with more than four subsystems (i.e., more than four series of loading complexes) have R at least 8 percent (but probably substantially more); multiple input schemes with four subsystems have R at least 4 percent (but probably substantially more). We do not have a complete proof that multiple input schemes with fewer than four subsystems have excessively high values of R (in comparison with (*d*) and (*e*)), but strongly believe that this is the case. Each subsystem of such a configuration would involve nine or more loading complexes in series.

3. Equipment Cost Data

Only those equipment costs which appeared likely to vary appreciably from one configuration to another were considered. The cost of coding the letters, for example, was neglected on the grounds that in a code sort system, all letters to be sorted must be coded once and only once, regardless of how many readings or sorts they undergo.

The two pieces of equipment whose costs were considered variable are the loading complexes and the modules of bins. Hereafter, the cost of bins for the sorted mail will be considered to include the costs of the corresponding conveyor, cart pockets, etc.

The following tentative cost estimates were given by a representative of the manufacturer:

$2,000/module of 30 bins, (3.1)

$125,060/first loading complex of a sorter, (3.2)

$50,000/each of the next 4 loading complexes of a sorter, (3.3)

$10,000/pair of end-pieces. (3.4)

Because (*a*) total equipment costs are somewhat lower than total personnel costs (see the three summands of (8.6), for example) and (*b*) the estimates (3.1) to (3.4) are only rough ones, it seemed permissible to be rather loose in our treatment of equip-

[3] These schemes are described as they arise later in the report

ment costs. First, the end-piece costs are small relative to other costs, and are difficult to handle because of the way in which the number of end-pieces depends on the particular placement of equipment. For these reasons, the end-piece costs have been neglected altogether. Second, there was apparently some doubt as to whether the extra loading complexes of a multi-input sorter can actually be used in the manner assumed in the estimate (3.3); for this reason we have disregarded (3.3) and taken the cost of every loading complex to be $125,000. Our numerical work for specific sorting schemes is fairly sensitive to these decisions, but the results of comparing different schemes are not.

In summary, the equipment costs *actually used* in what follows are

$$\$2{,}000/\text{module of 30 bins} \tag{3.5}$$
$$\$125{,}000/\text{loading complex} \tag{3.6}$$

These are initial costs; *yearly* costs are based on a *10-yr amortization period.*

4. Space Cost Data

The only space requirements taken into account [4] were those for the loading complexes themselves, for bins, and for the sweepers' work area around them:

$$8.75 \text{ ft} = \text{length of a loading complex.} \tag{4.1}$$
$$8.75 \text{ ft} = \text{length of a module of 30 bins.} \tag{4.2}$$
$$11.25 \text{ ft} = \text{width of a module, including sweepers' work space} \tag{4.3}$$

Using the tentative cost figure

$$\$2.20/\text{ft}^2/\text{yr} \tag{4.4}$$

suggested by L. Allison (NBS Electronic Instrumentation Section) we obtain $(8.75)\times(11.25)\times(2.2) \approx 220$, or

$$\$220/\text{loading complex/yr} \tag{4.5}$$

$$\$220/\text{module of 30 bins/yr.} \tag{4.6}$$

It is convenient to combine space and equipment costs. From (3.6) and (4.5), remembering the 10-yr amortization applying to (3.6), we have

$$\$12{,}500 + \$220 \approx \$12{,}700/\text{loading complex/yr.} \tag{4.7}$$

Similarly, from (3.5) and (4.6) we have $\$200 + \$220 = \$420$/module of 30 bins/yr, so that we have

$$\$14/\text{bin/yr.} \tag{4.8}$$

Formulas (4.7) and (4.8) are the ones used in our analysis.

5. Personnel Cost Data

The personnel which appears to vary appreciably from one configuration to another is composed of sweepers and pouchers. The tentative data used below were suggested by L. Allison.

Sweep operators remove sorted letters from the bins and either tie them out or place them in trays preceding a further sort. We take

$$60 \text{ letters} = \text{number of letters a sweeper can be expected to handle in one sweep,} \tag{5.1}$$

$$180 \text{ sweeps/hr} = \text{working rate expected of a sweeper.} \tag{5.2}$$

It is assumed that the sweeping is carried on in such away that:

$$\text{Each bin is swept at least once per hour,} \tag{5.3}$$

$$\text{if more than 60 letters/hr are expected in a particular bin, then (subject to over-rule by (5.3)) the bin is not swept when it contains fewer than 60 letters, and} \tag{5.4}$$

$$\text{no more than 20 sweeps (i.e., 1,200 letters)/hr is expected for any one bin.} \tag{5.5}$$

The rules (5.1) to (5.5) permit computation of the number of sweepers required, *if* it is known how many letters/hr are expected in each bin. This quantity can in turn be computed (as will be done in later sections) for any particular sorting scheme and mail distribution.

Pouching operators toss tied bundles of mail into pouches; their working rate is

$$540 \text{ sweeps pouched/hr.} \tag{5.6}$$

Comparing (5.2) and (5.6), we see that, for sorts which are immediately followed by pouching,

$$\text{number of pouchers} = 1/3 \times (\text{number of sweepers}). \tag{5.7}$$

As for sorts which are not followed by pouching (i.e., primary sorts to be followed by secondary sorts), we observe that the additional work involved in handling the trays and feeding the loading complexes for a secondary sort appears to be of the same order of magnitude as that of pouching a similar quantity of mail. Therefore, since we suppose the transfer between sorts to be nonautomatic, we will take the cost of the personnel needed for this transfer to be equal to that of a number of fictitious pouchers given by (5.7).[5]

The average salary figure used is

$$\$11{,}000/\text{man-position/yr,} \tag{5.8}$$

where a "man-position" requires more than one person because of the several shifts worked and the 7-day week involved.

6. Mail Distribution

The particular mail distribution assumed in our numerical work is a hypothetical one obtained by

[4] These gave the costs which appeared likely to vary appreciably from one configuration to another.

[5] In contrast, the personnel required for initial loading on the primary sort is nearly independent of the equipment configuration, and so is not considered.

modifying data of outgoing mail from Los Angeles.[6] It was decided that for a code sort machine, at least, all local, postage due, uncancelled mail, etc., could be removed at the coding station. The mail listed as go backs and residue (about 4%) was excluded from sorter input because of the difficulty of distributing it properly by destination. The mail distribution *to the sorting system*, therefore, consists approximately of what is left after these deletions, with the percentage of mail to each destination upgraded to bring the total to 100 percent, and the amount of mail to each destination upgraded to bring the total to

$$1{,}000{,}000 \text{ letters/hr.} \tag{6.1}$$

Our distribution involves

$$1{,}600 \text{ destinations.} \tag{6.2}$$

The details of the distribution are given in appendix A; they involve some inconsequential grouping in the "tail" of the distribution.

7. Cost Formula

The three variables in the cost formula are

$$L = \text{number of loading complexes,} \tag{7.1}$$
$$B = \text{number of bins, and} \tag{7.2}$$
$$S = \text{number of sweeper man-positions.} \tag{7.3}$$

The values of these variables can be found for any particular mail distribution and specified arrangement of equipment.

The total yearly cost C is given by

$$C = C_e + C_s + C_p,$$

where

C_e = yearly equipment cost,
C_s = yearly space cost, and
C_p = yearly personnel cost.

Using (4.7) and (4.8), we have

$$C_e + C_s = 12{,}700\, L + 14\, B;$$

using (5.7) and (5.8) we have

$$C_p = 11{,}000\ (S + \tfrac{1}{3}S) \approx 14{,}670\, S.$$

The *cost formula* we will use is therefore

$$C = 12{,}700\, L + 14\, B + 14{,}670\, S \tag{7.4}$$

where L, B, and S are defined above and

$$C = \text{variable dollar cost/yr.} \tag{7.5}$$

[6] N. C. Severo and A. E. Newman, A statistical chain ratio method for determining the distribution of mail by destination (to be published).

8. General Minimum Cost Estimate

We will first derive a theoretical minimum cost estimate which is *general* in the sense that it does not depend on the particular distribution by destinations of mail, but only on the *volume* of mail; i.e., on the fact that the sorting system must be able to handle

$$1{,}000{,}000 \text{ letters/hr.} \tag{8.1}$$

Using this and the fact that the maximum input rate is

$$36{,}000 \text{ letters/loading complex/hr,} \tag{8.2}$$

we see that *at least* $1{,}000{,}000/36{,}000 = 27^+$ loading complexes are required; since L is an integer,

$$L \geq 28. \tag{8.3}$$

By (5.5) each bin accounts for at most 1,200 letters/hr, so that in view of (8.1) the number of bins required is *at least*

$$1{,}000{,}000/1{,}200 = 833^+;$$

since bins come in 30-bin modules, B must be a multiple of 30, and so

$$B \geq 840. \tag{8.4}$$

By (5.1) each stack of letters contains at most 60 letters, so there must be *at least*

$$1{,}000{,}000/60 \quad \text{stacks/hr,}$$

and thus in view of (5.2) there must be *at least*

$$(1{,}000{,}000/60)/180 = 92^+ \text{ sweepers;}$$

thus

$$S \geq 93. \tag{8.5}$$

We now apply (8.3), (8.4), and (8.5) to (7.4), obtaining

$$C \geq (12{,}700) \times (28) + (14) \times (840) + (14{,}670) \times (93),$$

or, rounding off,

$$C \geq 356{,}000 + 12{,}000 + 1{,}364{,}000 \approx 1{,}732{,}000. \tag{8.6}$$

Thus we have proved that $1,732,000 is an (approximate) *minimum variable yearly total cost* for equipment, space, and personnel. The three summands in (8.6) refer, respectively, to costs associated with loading complexes, bins, and personnel. We emphasize that this is a *theoretical* minimum cost; no actual system can cost this little.

Next we will be a little more realistic (and a little less general), and use the following one fact about the particular mail distribution with which we deal: Each of the 700 least frequent destinations receive 60 or fewer letters/hr. On the one hand, they re-

quire at least 700 sweeps/hr and therefore yield at least 700 stacks/hr. On the other hand, these 700 destinations account for about 1.8 percent of the total mail, and thus for 18,000 letters/hr; if these letters were arranged into stacks of size 60 (as was assumed in deriving (8.5)), then

$$18{,}000/60 = 300 \text{ stacks/hr},$$

rather than 700, would result. Thus the argument leading to (8.5) underestimates the minimum possible number of stacks/hr by $700-300=400$, and thus (using (5.2)) underestimates S by $400/180 \approx 2$ sweepers. (8.5) should be replaced by

$$S \geq 95, \tag{8.7}$$

and the corresponding modification of (8.6), after rounding, is

$$C \geq 1{,}761{,}000,$$

so that we have a minimum cost estimate of

$$C_{\min} = \$1{,}761{,}000. \tag{8.8}$$

The estimate (8.8) is still too "general" to use as a yardstick, and we shall use instead a "detailed minimum cost estimate" (based on the detailed properties of the specific mail distribution) which will be derived in section 9. This care in choosing a yardstick might seen unnecessary, since a "yardstick" is only a unit of measurement whose choice cannot affect which of two proposed sorting systems appears less costly. The choice of yardstick does, however, affect our decisions as to what constitutes a *significant* difference in cost (either between two systems being compared or between a single system and a hypothetical "minimum-cost" system), and also as to what constitutes an *allowable* error in making simplifying approximations.

We shall use (8.8) primarily as an aid in calculating the costs of the various systems studied. If, using (8.3), (8.4), and (8.7), we define

$$\Delta L = L - 28 \tag{8.9}$$
$$\Delta B = B - 840 \tag{8.10}$$
$$\Delta S = S - 95, \tag{8.11}$$

then it turns out that a convenient way to calculate the cost C of a system is (*i*) to find these numbers ΔL, ΔB, and ΔS of "extra" loading complexes, bins, and sweepers, (*ii*) to calculate an "extra cost,"

$$\Delta C = C - C_{\min},$$

by $$\Delta C = (12{,}700)\ \Delta L + 14\ \Delta B + (14{,}670)\ \Delta S \tag{8.12}$$

(this formula follows from (7.4)), and (*iii*) to find C by

$$C = C_{\min} + (\Delta C). \tag{8.13}$$

This constitutes the "add up" method mentioned in section 1.

9. Detailed Minimum Cost Estimate

Our main goal in this section is to derive a minimum cost estimate which makes detailed use of the particular mail distribution we are studying. The methods used in this derivation will also turn out to be helpful in the analysis of some of the systems considered later in the report. Reading the first paragraph of section 12 may be helpful here.

All the systems studied later have the property that:

> For each destination, either all mail to that destination is sorted by destination on the primary, or all mail to that destination goes into residue on the primary and is then given a secondary sort. (9.1)

We shall therefore assume this property in (mentally) constructing a hypothetical "minimum-cost" system. Since this is to be a minimum-cost system, we can also suppose (to minimize the cost of residue bins) that:

> All residue bins are operating at their maximum capacity (see (5.5)) of 1,200 letters/hr. (9.2)

Once this assumption is made, it follows (in order to minimize loading complex costs) that:

> A letter which is to go into residue will be dropped into a residue bin before passing another loading complex. (9.3)

The method of estimating costs used in section 8 depended on counting the numbers of loading complexes, bins, and sweepers, and adding up the resultant costs. The approach used below is fundamentally different, in that it involves following every letter through the system and assigning an appropriate cost at each stage of its progress. We first consider a letter which enters the system through some loading complex $\mathscr{L}_1$ and then (without being dropped) passes a succeeding loading complex $\mathscr{L}_2$. This letter prevents $\mathscr{L}_2$ from working at its maximum capacity of 36,000 letters/hr; some other letter, which could otherwise have entered the system through $\mathscr{L}_2$, will now have to enter the system through some "extra" loading complex (in addition to the minimal 28 complexes known to be required by (8.3)). If we suppose (to minimize the number of extra complexes and thus the total cost of all extra complexes) that all extra complexes are operating at capacity, then our original letter, by passing $\mathscr{L}_2$ has created a requirement of "1/(36,000) of an extra loading complex." Since (by 7.4) each extra loading complex adds \$12,700 to the system, we are led to the following rule for use in estimating the cost of our hypothetical system:

> Assign a cost of $12{,}700/36{,}000 \approx .353$ whenever a letter passes a loading complex. (9.4)

Suppose now that it is possible to divide each extra loading complex into 36,000 parts, each able to handle 1 letter/hr, and that it is possible to add enough of these parts to each of the minimal 28 "nonextra" loading complexes required for the primary sort to ensure that these complexes operate at

their full 36,000 letters/hr capacity. Such a policy would tend to decrease bin costs and, indirectly, sweeper costs (since bins which previously received fewer than 60 letters/hr might now receive 60 or more; see (5.3)). We therefore assume, for our hypothetical minimum-cost system, that this policy has been adopted, so that *the primary sort involves 28 loading complexes* (each possibly augmented by parts of extra loading complexes) *all working at their maximum capacity*. If we also make the "minimum-cost" assumption that all loading complexes used for the secondary sort are operating at maximum capacity, then we have a situation in which:

All loading complexes operate at their maximum rate of 36,000 letters/hr. (9.5)

Thus any letter entering the primary or secondary sort uses 1/(36,000) of the services of the loading complex through which it enters, so that we have the following rule analogous to (9.4):

Assign a cost of .353 whenever a letter enters the primary or secondary sort. (9.6)

The essential assumptions made so far are (9.1), (9.2), (9.3), and (9.5). Before proceeding further, we point out that two of the (physically realizable) types of sorting configurations to be analyzed later actually *do* satisfy these assumptions. The simplex scheme (treated in sec. 10) satifies them exactly, obeying vacuously the conditions referring to residue; the optimal residue scheme found in section 12 satisfies them very nearly. Thus these schemes can be treated by the method developed below.[7] There is no need (see sec. 10) for such a detailed treatment of the simplex scheme. We shall treat the residue scheme, however; in fact, since the methods of this section are less obviously correct than is the "count the sweepers, bins, and loading complexes" approach, we will later analyze the residue scheme using each approach separately and verify that the same answer is obtained in both cases.

Returning to the analysis, we define

f_i=fraction of the mail which goes to the ith destination (9.7)

and consider whether or not letters to the ith destination should go into residue on the primary sort. For this purpose it is convenient to define

$$V_i=\begin{cases}1{,}000{,}000 f_i & \text{if } 1{,}000{,}000 f_i \leq 60, \\ 60 & \text{otherwise,}\end{cases} \quad (9.8)$$

and to note that (by (6.1))

1,000,000 f_i=number of letters/hr to the ith destination (9.9)

In addition, from (7.4) and (5.2) we are led to the rule:

Assign a (personnel) cost of 14,670/180=81.50 for each stack to be swept. (9.10)

If letters to the ith destination go into residue, then the total cost to be assigned to each such letter as it travels through the sorting system can be calculated in the following way. The letter enters the primary and secondary sorts, so (9.6) yields a loading complex cost of $2\times(.353)$. By (9.3), there is no contribution from (9.4). By (9.2) and (7.4), the letter should be "charged" 14/1,200 for its use of a residue bin. By (5.5) and (7.4), assuming $10^6 f_i \leq 1{,}200$, it should be charged $14/10^6 f_i$ for its use of a bin in the secondary sort. By (5.1) and (9.10) it should be charged 81.50/60 for the sweep of its residue bin, while by (5.1), (5.3) and (9.10) it should be charged $81.50/V_i$ for the sweep of its bin in the secondary sort. Adding together these partial costs and defining

$C(f_i)$=cost to be assigned to a letter to the ith destination, if such letters go into residue, (9.11)

we have

$$C(f_i)=(14/10^6 f_i)+(81.50/V_i)+2.076 \quad (\text{if } 10^6 f_i \leq 1{,}200). \quad (9.12)$$

Next we want to calculate the analogous cost if letters to the ith destination do *not* go into residue for a secondary sort. The primary sort of our hypothetical system now consists of 28 (possibly augmented) loading complexes in series, each followed by a row of bins (some of which may be residue bins); this series arrangement involves no loss in generality, since any other arrangement can be obtained as a special case.[8] We note first that

bins for the ith destination are spaced uniformly through the primary (except possibly at the end). (9.13)

To see why this is so, suppose for example that it is found that the first bins for the ith destination should be placed after the third loading complex. Then in view of (9.5), the situation regarding this destination is the same beginning with the fourth loading complex as it was beginning with the first complex, and for the same reasons as before we would find that the next bins for the destination should be placed as far after the fourth complex as the first bins were placed after the first complex. Thus we can define

n_i=number of loading complexes between successive appearances in the primary of bins for the ith destination. (9.14)

In view of (9.5), a letter to the ith destination is equally likely to pass 0, 1, 2, . . ., n_i-1 complexes, each with a probability $1/n_i$, so that the average cost contribution due to (9.4) is

[7] Actual schemes involving multiple input fail to obey (9.5); in treating these schemes by the methods developed below, one must take into account the actual input rates of the loading complexes.

[8] For example, to obtain as a special case (so far as cost factors are concerned) a multiple input system involving two series of 14 loading complexes each, we simply assume in our hypothetical system that all mail is dropped between the 14th and 15th complexes.

$$(.353)\times(0+1+2+\ldots+(n_i-1))/n_i=0.1765\,(n_i-1),$$

where the formula for the sum of an arithmetic progression has been used. There is a contribution of .353 due to (9.6). The system contains $28/n_i$ sets of bins for the ith destination, each set (disregarding fractional effects) receiving

$$1{,}000{,}000\,f_i/(28/n_i)\text{ letters/hr};$$

if we define

$$V_i'=\begin{cases}1{,}000{,}000\,f_i/(28/n_i) & \text{if this is } \leq 1{,}200,\\ 1{,}200 & \text{otherwise,}\end{cases}\qquad(9.15)$$

the letter should be charged (again ignoring fractional effects) $14/V_i'$ for its use of a bin. Similarly, if we define

$$V_i''=\begin{cases}1{,}000{,}000\,f_i/(28/n_i) & \text{if this is } \leq 60,\\ 60 & \text{otherwise,}\end{cases}\qquad(9.16)$$

then the letter should be charged (see (9.11)) $81.50/V_i''$ for the sweep of its bin. Adding together these partial costs and defining

$$C(f_i,\,n_i)=\text{cost to be assigned to a letter to the } i\text{th destination, if such letters are sorted directly on the primary,}\qquad(9.17)$$

we have

$$C(f_i,\,n_i)=0.1765\,n_i+(14/V_i')+(81.50/V_i'')+0.1765.\qquad(9.18)$$

We have now developed the formulas needed to investigate whether or not mail to a given destination should be put into residue. It is convenient to divide the destinations into three classes:

Class 1. $1{,}000{,}000\,f_i\leq 60$,
Class 2. $60<1{,}000{,}000\,f_i\leq 1{,}200$,
Class 3. $1{,}200<1{,}000{,}000\,f_i$.

In each class, the "into-residue" cost $C(f_i)$ should be compared with the "not-into-residue" cost $C(f_i,\,n_i)$ evaluated at that value of the system design parameter n_i which minimizes it.

Analysis of class 1. Since the first alternative of (9.8) holds, (9.12) yields

$$C(f_i)=2.076+(0.955\times 10^{-4})/f_i.\qquad(9.19)$$

Since the first alternatives of (9.15) and (9.16) apply, (9.18) yields

$$C(f_i,n_i)=0.1765\,n_i+(2.674\times 10^{-3})/(n_i f_i)+0.1765,\qquad(9.20)$$

so that $\partial C(f_i,\,n_i)/\partial n_i=0.1765-2.674\times 10^{-3}/(n_i^2 f_i)$,

from which it follows that $C(f_i,\,n_i)$ is a decreasing function of n_i for

$$n_i^2\leq 1.515\times 10^{-2}/f_i,$$

i.e., for

$$n_i\leq 0.1231/\sqrt{f_i},$$

and is increasing for higher values of n_i. Since $n_i\leq 28$, the minimizing value of n_i is

$$\begin{cases}0.1231/\sqrt{f_i} & \text{if this is } \leq 28,\\ 28 & \text{otherwise,}\end{cases}$$

i.e., is

$$\begin{cases}0.1231/\sqrt{f_i} & \text{if } f_i\geq 4.439\times 10^{-3},\\ 28 & \text{otherwise.}\end{cases}\qquad(9.21)$$

By (9.20) and (9.19)

$$C(f_i,\,28)=5.1185+(0.955\times 10^{-4})/f_i>C(f_i),$$

so that if the second alternative of (9.21) holds then mail to the ith destination should go into residue. If the first alternative holds in (9.21) then

$$C(f_i,n_i)_{\min}=C(f_i,\ \ 0.1231/\sqrt{f_i})=0.1765+(.04345/\sqrt{f_i}),$$

and (from (9.21) and the fact $1{,}000{,}000\,f_i\leq 60$)

$$4.439\times 10^{-3}\leq\sqrt{f_i}\leq 7.746\times 10^{-3},$$

$$129.1\leq x=(1/\sqrt{f_i})\leq 225.3.\qquad(9.22)$$

From (9.19) and our expression for $C(f_i,\ n_i)_{\min}$, we then have

$$C(f_i,n_i)_{\min}-C(f_i)=0.04345\,x-0.955\times 10^{-4}x^2-1.8955;$$

this function of x is increasing (i.e., its derivative is positive) in the range (9.22) and is >0 at $x=129.1$; thus it is positive throughout the range (9.22), which shows that even if the first alternative of (9.21) holds, mail to the ith destination should go into residue. Thus in our minimum-cost system, *all mail to class 1 destinations should go into residue.*

For our particular mail distribution (see app. A) the class 1 destinations are destinations 901 to 1,600, and these 700 destinations receive 18,080 letters/hr. These letters lead via (9.6) to a cost of

$$2\times(.353)\times 18{,}080$$

associated with loading complexes, of

$$14\times(18{,}080/1{,}200)+14\times 700$$

associated with residue bins and secondary bins, and of

$$81.50\times(18{,}080/60)+81.50\times 700$$

associated with sweeps of residue bins and secondary bins. Adding these, we find that in our minimum-

cost system class 1 destinations involve a total cost of approximately

$$\$104{,}400. \tag{9.23}$$

Analysis of class 2. Since the second alternative of (9.8) holds, (9.12) yields

$$C(f_i) = (14/10^6 f_i) + 3.434. \tag{9.24}$$

The first alternative of (9.15) holds, but either alternative of (9.16) may apply, so that (9.18) yields

$$C(f_i, n_i) = 0.1765 n_i + (2.674 \times 10^{-3})/(n_i f_i) + 0.1765 \quad (\text{if } n_i \leq 1680/10^6 f_i), \tag{9.25}$$

$$C(f_i, n_i) = 0.1765 n_i + (0.392 \times 10^{-3})/(n_i f_i) + 1.5348 \quad (\text{if } n_i > 1{,}680/10^6 f_i). \tag{9.26}$$

As in the analysis of class 1, the function given in (9.25) is a decreasing function of n_i for

$$n_i \leq 0.1231/\sqrt{f_i} \tag{9.27}$$

and is increasing for higher values of n_i. On the other hand, the function given by (9.26) is a decreasing function of n_i for

$$n_i \leq 0.04713/\sqrt{f_i} \tag{9.28}$$

and is increasing for higher values of n_i.

For further analysis it is convenient to divide class 2 into subclasses, depending on the relative positions of $n_i = 0.1231/\sqrt{f_i}$ (where the function given by (9.25) has its minimum), $n_i = 0.04713/\sqrt{f_i}$ (where the function given by (9.26) has its minimum), and $n_i = 1{,}680/10^6 f_i$ (where the formula for $C(f_i, n_i)$ changes from (9.25) to (9.26)). This subdivision leads to

Class 2a. $60 < 1{,}000{,}000\, f_i \leq 186.3$,
Class 2b. $186.3 < 1{,}000{,}000\, f_i \leq 1{,}200$,

corresponding respectively (subject to $60 < 10^6\, f_i \leq 1{,}200$) to

$$0.04713/\sqrt{f_i} < 0.1231/\sqrt{f_i} \leq 1{,}680/10^6 f_i, \tag{9.29a}$$

$$0.04713/\sqrt{f_i} \leq 1{,}680/10^6 f_i < 0.1231/\sqrt{f_i}. \tag{9.29b}$$

Analysis of class 2a. Here $C(f_i, n_i)_{\min}$ must arise either by setting $n_i = 0.1231/\sqrt{f_i}$ in (9.25), yielding

$$0.1765 + (0.04345/\sqrt{f_i}),$$

or by setting $n_i = 1{,}680/10^6 f_i$ in (9.26), yielding

$$1.7681 + (296.5/10^6 f_i).$$

It turns out that the first form is the smaller, so that, combining the first form with (9.24), we have

$$C(f_i, n_i)_{\min} - C(f_i) = (0.04345/\sqrt{f_i}) - (14/10^6 f_i) - 3.258. \tag{9.30}$$

This function of f_i is positive for

$$60 < 1{,}000{,}000\, f_i < 168.2, \tag{9.31}$$

which shows that mail to the ith destination *should* go into residue if (9.31) holds.

For our particular mail distribution, destinations 526 to 900 are the ones obeying (9.31); these 375 destinations receive 40,250 letters/hr, leading to a cost of

$$2 \times (.353) \times (40{,}250)$$

associated with loading complexes, of

$$12 \times (40{,}250/1{,}200) + 14 \times 375$$

associated with residue bins and secondary bins, and of

$$81.50 \times (40,\ 250/60) + 81.50 \times (40,\ 250/60)$$

associated with sweeps of residue bins and secondary bins. Adding these, we find that in our hypothetical minimum-cost system the mail to destinations obeying (9.31) has a total associated cost of approximately

$$\$143{,}500. \tag{9.32}$$

Before proceeding further, we will write out explicitly a principle which will be helpful in much of the following work. The correctness of the principle could be proved analytically in each situation in which we invoke it, but such proofs would be lengthy and repetitious, and so we content ourselves here with stating the principle (which is intuitively evident anyhow). Informally, the principle simply recognizes the fact that bins for destinations receiving a good deal of mail should appear rather frequently in the primary of our minimum-cost system (i.e., n_i will be small), bins for destinations receiving less mail should appear less frequently, (i.e., n_i will be larger), and there is a "threshhold t" such that the destinations whose mail is put into residue are precisely those which receive fewer than t letters/hr. Formally:

> Suppose it is best in our minimum-cost system that mail to the ith destination *not* go into residue, and suppose that $(n_i)_{\text{opt}}$ is the best value of n_i. Then for any other destination with $f_j \geq f_i$, it is best that mail to this destination also not go into residue, and furthermore $(n_j)_{\text{opt}} \leq (n_i)_{\text{opt}}$. (9.33)

Returning to the analysis, we now consider the class 2a destinations not obeying (9.31); i.e., those for which

$$168.2 \leq 1{,}000{,}000\, f_i \leq 186.3. \tag{9.34}$$

For these destinations, the function of f_i given in (9.30) is negative, but we cannot automatically con-

clude that mail to these destinations should not go into residue. The difficulty is that we treated n_i as a continuous variable in finding $C(f_i, n_i)_{\min}$, whereas in fact n_i must be a positive integer. We therefore define (for fixed f_i)

$$C(f_i, n_i)_{\min}^{(\text{int})} = \text{minimum of } C(f_i, n_i) \text{ for all positive integer values of } n_i; \tag{9.35}$$

the value of n_i yielding $C(f_i, n_i)_{\min}^{(\text{int})}$ will be one of the integers between which the n_i yielding $C(f_i, n_i)_{\min}$ lies.

We begin at the "top" of (9.34), with

$$1{,}000{,}000\, f_i = 186.3. \tag{9.36}$$

The value of n_i yielding $C(f_i, n_i)_{\min}$ is

$$n_i = 0.1231/\sqrt{f_i} = 9^+,$$

so that $(n_i)_{\text{opt}}$ is either 9 or 10. From (9.24) we have

$$C(f_i) = C(186.3 \times 10^{-6}) \approx 3.51.$$

Substituting $n_i = 9$ and $n_i = 10$ into the applicable choices of (9.25) and (9.26) (since $1{,}680/10^6 f_i = 9^+$, $n_i = 9$ goes into (9.25) and $n_i = 10$ into (9.26)), we obtain costs of approximately

$$3.34 \text{ for } n_i = 9, \qquad 3.51 \text{ for } n_i = 10,$$

so that $(n_i)_{\text{opt}} = 9$ and mail to destinations obeying (9.36) should not go into residue in our hypothetical minimum-cost system. We note for future reference that, by (9.33), all destinations in classes 2b and 3 should not have their mail sent into residue, and that throughout these classes $(n_i)_{\text{opt}} \geq 9$.

We now continue "down" through (9.34), considering lower values of f_i. Throughout (9.34),

$$n_i = 0.1231/\sqrt{f_i} = 9^+$$

so that $(n_i)_{\text{opt}}$ is either 9 or 10; also, throughout (9.34) we have $1{,}680/10^6 f_i = 9^+$, so that $n_i = 9$ must be substituted into (9.25), $n_i = 10$ must be substituted into (9.26), and the results compared with the result of (9.24). Since this yields

$$C(f_i, 9) = 1.7650 + (297.1/10^6 f_i),$$
$$C(f_i, 10) = 3.2998 + (39.2/10^6 f_i),$$

we find (using (9.24) also) that mail should go into residue for

$$168.2 \leq 1{,}000{,}000\, f_i \leq 169.6 \tag{9.37}$$

and should not go into residue for

$$169.6 < 1{,}000{,}000\, f_i \leq 186.3, \tag{9.38}$$

and also that

$$(n_i)_{\text{opt}} = 9 \qquad \text{in (9.38)}. \tag{9.39}$$

For our particular mail distribution, there are no destinations in the narrow range (9.37). The cost associated with destinations obeying (9.38) will not be found here, since it can more conveniently be combined with the cost found for class 2b.

Analysis of class 2b. We know that mail to these destinations does not go into residue. Also, we know that at the "bottom" of class 2b (i.e., $10^6 f_i = 186.3$) we have $(n_i)_{\text{opt}} = 9$, and from (9.33) we can expect that as we search "upward" through class 2b (i.e., examine successively larger values of f_i) we will reach a point where $(n_i)_{\text{opt}}$ changes from 9 to 8. Our immediate aim is to find this turnover point.

We find that for $186.3 \leq 10^6 f_i \leq 186.7$, $C(f_i, 8)$ and $C(f_i, 9)$ are both given by (9.25), so that

$$C(f_i, 9) = 1.7650 + (297.1/10^6 f_i)$$
$$C(f_i, 8) = 1.5885 + (334.3/10^6 f_i) \tag{9.40}$$

Throughout the range $C(f_i, 9)$ is smaller, so that

$$(n_i)_{\text{opt}} = 9 \qquad \text{for} \qquad 186.3 \leq 10^6 f_i \leq 186.7.$$

For $186.7 < 10^6 f_i \leq 210$, $C(f_i, 8)$ is still given by (9.25) but $C(f_i, 9)$ is now given by (9.26), so that

$$C(f_i, 9) = 3.1233 + (43.56/10^6 f_i). \tag{9.41}$$

Comparing (9.40) and (9.41), we find that

$$(n_i)_{\text{opt}} = 9 \qquad \text{for} \qquad 186.7 < 10^6 f_i \leq 189.4$$

but

$$(n_i)_{\text{opt}} = 8$$

for values of $10^6 f_i$ immediately above 189.4.

We can continue to search upwards through class 2b, looking next for the turnover from $(n_i)_{\text{opt}} = 8$ to $(n_i)_{\text{opt}} = 7$, etc. The results of this search are given in table 1 which includes the class 2a destinations obeying (9.38). The quantity $10^6 f_i$ has been rounded.

The first three columns of table 1 are independent of the particular mail distribution. The first, third, fourth, and fifth columns can be used to derive a cost figure in the following way: For each row, the fourth entry shows how many destinations are involved and the first entry enables us to find how many bins to each destination appear in our minimum-cost system. If (9.25) holds, then these bins are each swept once an hour, whereas if (9.26) holds then the number of sweeps involved is the same as that used in deriving (8.7); we can use the fifth entries of the rows involving (9.25) to find the numbers of "extra sweepers" involved (see (8.11)) and then multiply this contribution to ΔS by 14,670 in accordance with (8.12). Finally, we can use the first and fifth columns, together with (9.4) and (9.6), to assign a loading complex cost to each row. The result[9] is a total cost of about

[9] Different methods of associating integral numbers of bins with values of n_i which are not divisors of 28 were tested and found to change (9.42) only negligibly.

$401,200 (9.42)

associated with class 2b destinations and the class 2a destinations obeying (9.38).

TABLE 1. *Optimizing n_i in class 2b*

$(n_i)_{opt}$	$10^6 f_i$	Formula	Destinations	Letters/hr
9	170 to 187	(9.25)	491 to 525	5,950
9	188 to 189	(9.26)		
8	190 to 209	(9.25)	451 to 490	7,700
8	210 to 217	(9.26)	441 to 450	2,100
7	218 to 239	(9.25)	421 to 440	4,500
7	240 to 254	(9.26)	401 to 420	4,900
6	255 to 279	(9.25)	381 to 400	5,300
6	280 to 305	(9.26)	351 to 380	8,800
5	306 to 335	(9.25)	341 to 350	3,200
5	336 to 384	(9.26)	311 to 340	10,800
4	385 to 419	(9.25)	300 to 310	4,510
4	420 to 516	(9.26)	265 to 299	16,120
3	517 to 559	(9.25)	253 to 264	6,300
3	560 to 786	(9.26)	203 to 252	32,720
2	787 to 839	(9.25)	200 to 202	2,420
2	840 to 1,200	(9.26)	154 to 199	46,400
Total number of letters/hr				161,720

Analysis of class 3. Here the different possible alternatives in (9.15) and (9.16) lead to three possible formulas for $C(f_i, n_i)$:

$$C(f_i, n_i)=0.1765n_i+(2.674\times10^{-3})/(n_if_i)+0.1765 \quad (\text{if } n_i\leq1{,}680/10^6f_i), \tag{9.43}$$

$$C(f_i, n_i)=0.1765n_i+(0.392\times10^{-3})/(n_if_i)+1.5348 \quad (\text{if } 1{,}680/10^6f_i<n_i\leq33{,}600/10^6f_i), \tag{9.44}$$

$$C(f_i, n_i)=0.1765n_i+1.546 \quad (\text{if } n_i>33{,}600/10^6f_i). \tag{9.45}$$

We can work "upwards" through class 3 just as we did in class 2b, beginning with $(n_i)_{opt}=2$; the only new complication is that possible transitions to (9.45) must be allowed for. It turns out that the change from $(n_i)_{opt}=2$ to $(n_i)_{opt}=1$ occurs for $10^6f_i=1{,}614$ (i.e., at destination 119 in our particular mail distribution) and that the cost associated with class 3 destinations is

$$\$1{,}397{,}000. \tag{9.46}$$

Finally, we add up (9.23), (9.32), (9.42), and (9.46) to obtain (after rounding off) a total of

$$C^*_{min}=\$2{,}046{,}000 \tag{9.47}$$

as the approximate cost of our hypothetical minimum-cost system. This is then our theoretical minimum cost, to be used as a yardstick in dealing with proposed sorting configurations. It is significantly larger than (8.8).

The general approach used in this section constitutes the "follow through" method mentioned in section 1.

10. Simplex Scheme

The first sorting scheme we examine is also the simplest. The sorting system consists of a number of essentially independent subsystems, each consisting of a single loading complex followed by 1,600 bins, one for each of the 1,600 destinations.

The argument used to derive (8.3) shows that for this system $L=28$, so that

$$\Delta L=0; \tag{10.1}$$

thus there are 28 subsystems, so that

$$\Delta B=(28)\times(1{,}600)-840\approx44{,}000. \tag{10.2}$$

Assuming the 1,000,000 letters/hr divide equally (on the average) among the 28 subsystems, we find that a bin corresponding to one of the 114 most frequent destinations will receive 60 or more letters/hr, whereas a bin corresponding to one of the $(1{,}600-114)=1486$ least frequent will receive fewer than 60. There are

$$1{,}486\times28\approx41{,}600$$

bins in the system which correspond to these last 1,486 destinations, and by (5.3) these bins give rise to 41,600 stacks/hr. These last 1,486 destinations receive only 27.6 percent of the total mail (276,000 letters/hr) and so, *if* all stacks consisted of 60 letters (the basis on which (8.5) was derived), they would give rise to only

$$276{,}000/60=4{,}600 \text{ stacks/hr.}$$

Thus we have $41{,}600-4{,}600=37{,}000$ extra stacks/hr, leading via (5.2) to

$$37{,}000/180\approx205$$

additional sweepers, of which two were taken into account in passing from (8.5) to (8.7). So

$$\Delta S=203 \tag{10.3}$$

and by (8.12)

$$\Delta C=(14)\times(44{,}000)+(14{,}670)\times(203)\approx3{,}594{,}000, \tag{10.4}$$

so that, by (8.8) and (8.13),

$$C=C_{min}+(\Delta C)=5{,}355{,}000,$$

and by (9.47),

$$(C-C^*_{min})/C^*_{min}\approx162\%.$$

Thus on a cost basis the simplex scheme should definitely be rejected. The excessive cost comes primarily from personnel, i.e., from the second summand in (10.4).

11. Christmas Tree Scheme

This might also be called the "square root" scheme. In our case the square root of the number

of destinations is $\sqrt{1,600}=40$. The system consists of a number of subsystems, each containing a single loading complex and 40 bins.[10] The 1,600 destinations are divided into 40 groups of 40 destinations each; the primary sorting of mail is done by group, and the mail to each group is then given a secondary sort by individual destinations.

The key to the analysis is the fact that each piece of mail is dealt with twice. By applying (8.3), (8.4), and (8.5) to both the primary and the secondary sorts, and adding the results, we have

$$L \geq 56, \qquad B \geq 1,680, \qquad S \geq 186,$$

so that

$$\Delta L \geq 28, \qquad \Delta B \geq 840, \qquad \Delta S \geq 91$$

and by (8.12)

$$\Delta C \geq 1,702,000,$$

so that by (8.8) and (8.13)

$$C \geq 3,463,000$$

and

$$(C - C^*_{\min})/C^*_{\min} > 69\%.$$

On this basis we can reject the Christmas Tree Scheme. Again personnel costs are the major factor.

12. Residue Scheme

In this system the primary sort consists of sorting letters directly to those destinations which receive a relatively large fraction of the mail, while dropping all letters to the less "frequent" destinations into a relatively small number of residue bins, which are then given a secondary sort by destination. The purpose of this maneuver is to avoid having a large number of sweeps (corresponding to infrequent destinations) resulting in small stacks; such sweeps lead to excessive personnel costs.

The system is determined by stating definitely which destinations are to be considered "infrequent" (so that mail to them goes into residue) and which are to be considered "frequent." We therefore define a system design parameter

t=threshold; the ith destination is frequent if $10^6 f_i \geq t$, infrequent if $10^6 f_i < t$; (12.1)

the value of t is to be chosen so as to minimize the cost of the system. As before,

f_i=fraction of mail which goes to the ith destination. (12.2)

We shall determine the optimal value of t by two methods (in order to check their agreement). First we apply the "follow through" approach of section 9. The residue scheme certainly satisfies (9.1) and (9.3); if well-designed, it will very nearly satisfy (9.2) and (9.5) as well. We will therefore proceed as in section 9. The number of letters/hr to the ith destination in each of the 28 primary subsystems is

$$1,000,000\, f_i/28,$$

and if this quantity is >60 then by (5.4) there would be no reason to put mail to the ith destination into residue. Thus the only destinations about which there is any question are those for which

$$(1,000,000\, f_i/28) \leq 60.$$

We can split these destinations into classes, according as

$$1,000,000\, f_i \leq 60, \tag{12.3a}$$

or

$$60 < 1,000,000\, f_i \leq 1,200, \tag{12.3b}$$

or

$$1,200 < 1,000,000\, f_i \leq 1,680. \tag{12.3c}$$

For destinations obeying (12.3a), (9.12) yields

$$C(f_i) = 2.076 + (95.50/10^6 f_i), \tag{12.4a}$$

while (9.18), since $n_i=1$ in the residue scheme, yields

$$C(f_i, 1) = 0.353 + (2674/10^6 f_i). \tag{12.5a}$$

Similarly, for destinations obeying (12.3b) we have

$$C(f_i) = 3.434 + (14/10^6 f_i) \tag{12.4b}$$

$$C(f_i, 1) = 0.353 + (2,674/10^6 f_i). \tag{12.5b}$$

For destinations obeying (12.3c), it is not clear in view of (5.5) what a letter going into residue should be "charged" for the sweep of its secondary bin; it should be *at least* 14/1,200, however, so that (see (9.25)) we have

$$C(f_i) < (14/1,200) + 3.434 = 3.451, \tag{12.4c}$$

and

$$C(f_i, 1) = 1.7113 + (392/10^6 f_i). \tag{12.5c}$$

We find that $C(f_i) - C(f_i, 1)$ is negative (i.e., it is less expensive to put mail to the ith destination into residue) for $10^6 f_i \leq 863$ and is positive for $10^6 f_i \geq 864$. Thus the optimal t is

$$t_{\text{opt}} = 864, \tag{12.6}$$

which for our particular mail distribution corresponds to destination 196.

Next we apply the "add up" method based on (8.12). First we define two relevant quantities, depending on t, by

N_t=number of frequent destinations,
V_t=number of letters/hr to frequent destinations,

so that (since our situation involves in all 1,000,000

[10] The cost estimate would be even higher if we took into account the indivisibility of the 30-bin modules.

letters/hr and 1,600 destinations).

$1{,}600-N_t$=the number of infrequent destinations,
$1{,}000{,}000-V_t$=number of letters/hr to infrequent destinations.

The number ΔL of extra loader complexes needed for the secondary sort is given (using (8.2)) by

$$\Delta L \approx (1{,}000{,}000-V_t)/36{,}000; \qquad (12.7)$$

more precisely, ΔL is the integer next above the right side of (12.7). The infrequent destinations are divided into ΔL *groups* whose total expected hourly mail volumes are approximately equal.

The sorting system contains $28+(\Delta L)$ subsystems, each consisting of a loading complex followed by a number of bins. Each of the 28 subsystems needed for the primary sort contains N_t bins for the frequent destinations and one residue bin[11] for each of the ΔL groups of infrequent destinations. Each of the ΔL subsystems used for the secondary sort, sorts the entire primary residue of some one of the groups; thus these ΔL subsystems together contain one bin[11] for each infrequent destination, and thus contain $1{,}600-N_t$ bins in all. Therefore

$$\begin{aligned} B &= 28(N_t+(\Delta L))+(1{,}600-N_t) \\ &= 28(\Delta L)+27N_t+1{,}600, \end{aligned}$$

$$\Delta B = B-840 = 28(\Delta L)+27N_t+760. \qquad (12.8)$$

Having found ΔL and ΔB, we must find ΔS. The frequent destinations receiving 60 or more letters/hr on each of the 28 primary subsystems, require no more sweeps than was assumed in calculating the minimum cost. The same holds for *secondary* sweeps of mail to the infrequent destinations. The *extra* sweeps therefore arise (*a*) from sweeping the residue bins and (*b*) from the primary sweeps of frequent destinations receiving fewer than 60 letters/hr on each of the 28 primary-sorting subsystems; the frequent destinations described in (*b*) are those receiving fewer than $(28)\times(60)=1{,}680$ letters/hr in all, and thus those whose mail frequencies are less than 0.168 percent. In our actual mail distribution, 114 destinations have frequencies of 0.168 percent or more, and these destinations account for 72.40 percent of the total mail. Thus

N_t-114=number of destinations of the type (b), (12.9)

$V_t-724{,}000$=hourly volume of mail to these destinations. (12.10)

The destinations described in (*b*) thus lead to approximately

$$28\ (N_t-114)/180 \text{ sweepers}$$

of which $((V_t-724{,}000)/60)/180$ were accounted for in calculating (8.8). The number of extra sweepers due to source (*b*) is therefore approximately

$$(1{,}680N_t-V_t+532{,}500)/(60)(180).$$

The residue bins together receive $1{,}000{,}000-V_t$ letters/hr and each receives more than 60 letters/hr, so that the number of extra sweepers due to source (*a*) is approximately

$$(1{,}000{,}000-V_t)/(60)\times(180);$$

adding the last two expressions gives

$$\Delta S \approx (1{,}532{,}500-2V_t+1{,}680N_t)/(60)(180). \qquad (12.11)$$

We now substitute (12.8) and (12.11) into (8.12) obtaining

$$\Delta C \approx 13{,}100(\Delta L)+2{,}660N_t-2.717V_t+2{,}092{,}500. \qquad (12.12)$$

The error involved in using (12.7) is at most unity, leading to an error of at most 13,100 in (12.12) and thus to an error of less than 1 percent in $(C-C^*_{\min})/C^*_{\min}$. We therefore substitute (12.7) into (12.12) getting

$$\Delta C \approx 2{,}660\ N_t-3.081\ V_t+2{,}456{,}400. \qquad (12.13)$$

To minimize ΔC quickly (we omit the rigorous justification of the following method), equate the differential of the right side of (12.13) to zero:

$$2{,}660(dN_t)-3.081(dV_t)=0.$$

Since N_t increases in steps of size 1 (each step involving shifting the status of one destination from "infrequent" to "frequent"), we set $dN_t=1$ and obtain

$$dV_t=(2{,}660)/3.081\approx 863 \text{ letters/hr.} \qquad (12.14)$$

Since the increment in V_t due to one extra "frequent" destination is simply the expected hourly volume of mail to that one destination, (12.11) tells us that for the approximate minimization of ΔC, the last of the "frequent" destinations should obey $10^6 f_t=863$, so that (see (12.1))

$$t_{\text{opt}}=863, \qquad (12.15)$$

in near-perfect agreement with the result (12.6) obtained by the "follow through" method. For our particular mail distribution this "cutoff" occurs between destinations 196 and 197, and after obtaining the value of V_t corresponding to $N_t=196$ from appendix A and substituting into (12.14) we find (rounding) that

$$(\Delta C)_{\min}=440{,}000$$

[11] Some extra bins may be required in order to satisfy (5.5), but their cost is negligible.

so that $C=2{,}201{,}000$

and $$(C-C^*_{\min})/C^*_{\min}\approx 8\%. \tag{12.16}$$

Using (12.11) we find that the cost of personnel is again the main cost factor.

The optimum given by (12.15) (i.e., given by a "cutoff" between destinations 196 and 197) is not a very sharp one; the cutoff can occur as low as about destination 165 or as high as destination 225 without raising $(C-C^*_{\min})/C^*_{\min}$ to more than 0.5 percent above its minimum.

13. Multiple-Input Schemes

In these schemes, the sorting system consists of a number of identical subsystems, each receiving its input from a number of loading complexes arranged in series. More precisely, each subsystem consists of a first loading complex followed by a first row of bins for some (but not all) destinations, then a second loading complex[12] followed by a second row of bins for some destinations (not necessarily the same ones as in the first row), . . ., and finally a last loading complex followed by a last row of bins for all 1,600 destinations.

Consider some particular sorting system of this type; call it *system 1*, and let

$$M=\text{number of subsystems of system 1.} \tag{13.1}$$

Let *system 2* be obtained from system 1 by replacing, with residue bins, the last 700 bins of the last row of each subsystem (these bins correspond to destinations 900 to 1600, which for our mail distribution are the ones receiving fewer than 60 letters/hr). The residue then requires a secondary sort. Let

$$R=\text{number of residue bins/subsystem for system 2.} \tag{13.2}$$

We will compare the costs of system 1 and system 2; let L_1, B_1, S_1, C_1, L_2, B_2, S_2, C_2 denote the respective values of L, B, S, C, for the two systems.

Destinations 900 to 1,600 receive about 1.8 percent of the mail, or 18,000 letters/hr; this is less than the 36,000 letters/hr capacity of a loading complex; and so only one complex is needed for the secondary sort:

$$L_2-L_1=1. \tag{13.3}$$

The 700 M bins used for destinations 900 to 1,600 in the last rows of the subsystems of system 1 are replaced in system 2 by RM residue bins plus 700 bins for the secondary sort, and so

$$B_1-B_2=700M-(RM+700). \tag{13.4}$$

The above-mentioned 700 M bins for system 1 each received fewer than 60 letters/hr, and so together required 700 M sweeps/hr. For system 2, the 700 secondary bins together require 700 sweeps/hr. As for the RM residue bins, R is chosen as small as possible, so that each residue bin (except possibly for the last one on each subsystem) receives at least 60 letters/hr; the residue bins receive 18,000 letters/hr and thus require approximately

$$18{,}000/60\approx 300 \text{ sweeps/hr.}$$

Thus we have, using (5.2),

$$S_1-S_2\approx(700M-(300+700))/180. \tag{13.5}$$

The analog of (8.12) which applies to our situation is

$$C_1-C_2=12{,}700(L_1-L_2)+14(B_1-B_2)+14{,}670(S_1-S_2), \tag{13.6}$$

and so, by (13.3), (13.4), and (13.5) we have

$$C_1-C_2\approx 67{,}000M-14RM-104{,}000. \tag{13.7}$$

Since R is chosen as small as (5.5) permits, we have

$$RM\approx 19{,}400/1{,}200\approx 16,$$

and so (13.7) can be rewritten as

$$C_1-C_2\approx 67{,}000M-104{,}000. \tag{13.8}$$

One rather sweeping conclusion which can be drawn from (13.8) is that *no multiple input scheme with more than one subsystem* (i.e., with $M>1$) *can be optimal*. To prove this, take the system in question as the "system 1" of the above discussion. Since $M>1$, (13.8) shows that $C_1-C_2>0$ and thus that $C_2<C_1$; since system 2 costs less than system 1, the latter cannot be optimal. We could use this argument to eliminate multiple input schemes with $M>1$ from further discussion, *if* we were going later to examine the "systems 2." Unfortunately, the analysis of such systems, which combine elements of the multiple input and residue schemes, appears too complicated to be attempted here (the difficulties of even relatively simple systems of this type will become apparent in sec. 14 and app. C). We can, however, use (13.8) to eliminate all multiple input systems (i.e., all "systems 1") *with* $M\geq 7$. To do this, we note first that $C_2>C_{\min}$, so that

$$\Delta C_1=C_1-C_{\min}>C_1-C_2;$$

together with (13.8) and $M\geq 7$, this implies that

$$\Delta C_1>C_1-C_2\approx 67{,}000\,M-104{,}000\geq 365{,}000,$$

so that system 1 is at best negligibly less costly than (actually *more* costly; see footnote 13, p. 96) the

[12] Some of the letters inserted by the first loading complex (namely, letters to those destinations for which bins were not provided in the first row of bins) will still be on the conveyor as it passes the second loading complex. Thus the second loading complex of each subsystem will *not* operate at its full rate (36,000 letters/hr) and the same holds for the third, fourth, etc. loading complexes of each subsystem.

system to be found in section 14, for which (see (14.24))

$$C \approx 366{,}000 \tag{13.9}$$

Now we want to examine multiple-input schemes with $M \leq 6$ (i.e., with six or fewer subsystems). It is convenient to define

$$N_M = \text{number of destinations receiving fewer than 60M letters/hr,} \tag{13.10}$$

$$W_M = \text{number of sweeps associated with these destinations in calculating } C^*_{\min}. \tag{13.11}$$

Each such destination receives fewer than 60 letters/hr (and thus requires at least one sweep/hr) for each subsystem, so that together these destinations require *at least* MN_M sweeps/hr for the entire system. Thus

$$\Delta S \geq (MN_M - W_M)/180. \tag{13.12}$$

For $M=6$ we find $MN_M \approx 7{,}620$, $W_M \approx 2{,}140$, so that, by (8.12),

$$\Delta C \geq 14{,}670 \times 5{,}480/180 \approx 447{,}000 \quad (\text{for } M=6). \tag{13.13}$$

By comparison with (13.9), *systems with $M=6$ are eliminated.*[13] For $M=5$ we find $MN_M \approx 6{,}150$, $W_M \approx 1{,}925$, as well as

$$\Delta B = B - 840 \geq 5 \times 1{,}600 - 840 = 7{,}160; \tag{13.14}$$

thus by (8.12),

$$\Delta C \geq 14{,}670 \times 4{,}225/180 + 14 \times 7{,}160 \approx 445{,}000 \quad (\text{for } M=5). \tag{13.15}$$

By comparison with (13.9), *systems with $M=5$ are eliminated.*

These relatively simple arguments are not adequate to deal with the schemes with $M \leq 4$. Consider first the situation $M=4$. Here $MN_M \approx 4{,}720$, $W_M \approx 1{,}710$ so by (13.12)

$$\Delta S \geq 3{,}010/180. \tag{13.16}$$

Two cases are possible. *If* each subsystem contains eight or more loading complexes, then

$$\Delta L = L - 28 \geq 4 \times 8 - 28 = 4, \tag{13.17}$$

$$\Delta B = B - 840 \geq 4 \times 1{,}600 - 840 = 5{,}560, \tag{13.18}$$

and by (13.16), (13.17), (13.18), and (8.12), we obtain

$$\Delta C \geq 374{,}000,$$

so that in this case the system can be eliminated by comparison with (13.9). The other possible alternative is that each subsystem contains no more than seven loading complexes. Seven loading complexes can handle at most

$$7 \times 36{,}000 = 252{,}000 \text{ letters/hr,}$$

and each of the four subsystems must handle

$$1{,}000{,}000/4 = 250{,}000 \text{ letters/hr,}$$

so seven complexes/subsystem are needed, all working close to capacity; i.e., the input of each loading complex in a subsystem must be nearly all dropped in the following row of bins in order for the next loading complex to have a nearly empty conveyor belt. This clearly requires at least 200 bins after each of the first 6 complexes in each subsystem; the seventh loader is followed by 1,600 bins (one for each destination) so that

$$\Delta B = B - 840 \geq 4 \times ((6 \times 200) + 1{,}600) - 840 = 10{,}360; \tag{13.19}$$

from (13.16), (13.19) and (8.12) we obtain

$$C \geq 390{,}000,$$

so that the system is eliminated by comparison with (13.9). Thus *systems with $M=4$ are eliminated.*

We have not been able, within the limits of time and effort reasonably assignable to this particular point,[14] to devise a mathematical proof that multiple input systems with $M=1, 2,$ or 3 can be eliminated because of excessive cost. (The difficulties encountered, and the reason for their occurrence for *small* values of M, are discussed in the next paragraph.) Nevertheless, we strongly believe that such systems are excessively costly. This belief is based on auxiliary calculations which will not be reproduced here, and also in general on our experience with the other parts of this study. Roughly, the situation is this: In each subsystem, most of the mail must be dropped out fairly soon after it enters the subsystem, for otherwise the input from many of the loading complexes would be cut substantially below 36,000 letters/hr, so that a large number of loading complexes (whose cost would be excessive) would be required to pass the required 1,000,000 letters/hour into the system. Such an early dropout of most mail, however, would require a large number of bins, many of which would receive fewer than 60 letters/hr and thus (see (5.3) and (5.4)) involve "inefficient sweeping"; the cost of bins and sweepers (especially the latter) would then be excessive.

The difficulties encountered for small values of M (i.e., for systems with a small number of subsystems) stem from the fact that in such cases each subsystem must contain a relatively *large* number of loading complexes; for $M=3$, for example, each subsystem contains at least 9 such complexes, while for $M=1$ we have a single long subsystem with at least 28 loading complexes in series. The first cause of difficulty is that the formulas involved in the analysis of a subsystem with n loading complexes

[13] The same argument, if applied to $M=7$, yields $\Delta C \geq 552{,}000$, a much stronger result than the one found earlier (the display first above (13.9)).

[14] Especially since the configuration found in section 14 is so nearly optimal.

become more and more complicated as n increases. (See Appendix B for more details.) For $n \geq 9$, for example (i.e., for $M \leq 3$), it is still possible to use these formulas in making calculations for any one given subsystem, but is extremely difficult to use them in comparing a large number of possible subsystems; this latter problem is of course, the one which actually arises in our work. The second cause of difficulty is the enormous number of systems to be compared, when the value of n is high (i.e., when M is low); this number is in fact

$$(n-1)^{(2^d)} \qquad (d = \text{number of destinations}),$$

and since $d = 1{,}600$ it is clear that even if 99 percent of the possible systems could be eliminated on some common-sense grounds, the number remaining for analysis (if n is moderately large) would still be astronomical.

14. Multiple Input and Residue: Double Loading

The "multiple input and residue" schemes, combinations of two of the proposed schemes which we have analyzed earlier, are like multiple input schemes except that some of the bins may be assigned to residue; this residue then requires a secondary sort. We therefore speak of "primary subsystems" and "secondary subsystems."

There is a great variety of subclasses of this type of scheme, and within any one subclass the analysis required to determine the optimal choice of the relevant parameters appears to be quite difficult. We will make a detailed analysis only of a relatively simple subclass; the result turns out to be a system so nearly optimal (see (14.25)) that investigation of more complicated systems is clearly not worthwhile.

We consider "multiple input and residue" schemes which are determined by two parameters, j and k (with $j<k$), in the following way: Each primary subsystem contains two loading complexes. The first complex of each primary subsystem is followed by bins for each of

the "type 1" destinations $1, 2, \ldots, j$, (14.1)

and also by *residue* bins for

the "type 3" destinations $k+1, k+2, \ldots, 1{,}600$, (14.2)

Thus that part of the first complex's input consisting of mail to

the "type 2" destinations $j+1, j+2, \ldots, k$ (14.3)

does not get sorted until after the conveyor passes the second complex. The second complex of each primary subsystem is followed by bins for type 1 and type 2 destinations, and residue bins for the type 3 destinations. We wish to choose j and k so as to minimize the cost of the system.

We will use the notation

$$f_i = \text{fraction of mail to destination } i; \quad (14.4)$$

the destinations are so ordered that

$$f_1 \geq f_2 \geq f_3 \geq \ldots \geq f_{1,600}.$$

We also set

$$F_i = f_1 + f_2 + \ldots + f_i = \text{fraction of mail to the first } i \text{ destinations.} \quad (14.5)$$

The input to the first loading complex of each primary subsystem is the usual 36,000 letters/hr, but only a fraction $F_j + (1-F_k)$ of these (corresponding to types 1 and 3) get dropped before the second complex, so that the input to the second complex is

$$36{,}000\ (1+F_j-F_k) \text{ letters/hr}$$

and the input to each primary system is

$$36{,}000 + 36{,}000\ (1+F_j-F_k) = 36{,}000\ (2+F_j-F_k) \text{ letters/hr.}$$

Therefore the required *number of primary subsystems* is about

$$1{,}000{,}000/36{,}000\ (2+F_j-F_k) \approx 28/(2+F_j-F_k), \quad (14.6)$$

so that the number of loading complexes for the primary subsystems is about

$$56/(2+F_j-F_k),$$

while the residue of $1{,}000{,}000(1-F_k)$ letters/hr requires

$$1{,}000{,}000\ (1-F_k)/36{,}000 \approx 28\ (1-F_k) \quad (14.7)$$

loading complexes for secondary sort. Thus

$$L \approx (56/(2+F_j-F_k)) + 28(1-F_k),$$

or
$$\Delta L = L - 28 \approx (56/(2+F_j-F_k)) - 28F_k \quad (14.8)$$

The type 3 destinations are divided into *groups* as for the residue scheme (sec. 12); the number of these groups is given by (14.7). The secondary subsystems together contain one bin for each type 3 destination, or $1{,}600-k$ bins in all. Each primary subsystem contains $j+k$ bins for separate destinations (j bins before the second loading complex, k bins after it), and also two residue bins for each group (one before the second complex, one after it). Hence, using (14.6) and (14.7),

$$B \approx (1{,}600-k)+(28/(2+F_j-F_k)) \times (j+k+56\ (1-F_k)),$$

so

$$\Delta B = B-840 \approx (760-k) +28\ (j+k+56\ (1-F_k))/(2+F_j-F_k). \quad (14.9)$$

Having found ΔL and ΔB, we still have the more complicated task of finding ΔS. The sweeps of bins of the secondary sort require no *extra* sweepers. The extra sweepers are required (*a*) in sweeping the residue bins,[15] (*b*) in sweeping those bins (if any) *before* the second loading complex which receive fewer than 60 letters/hr, and (*c*) in sweeping those bins *after* the second complex which receive fewer than 60 letters/hr. There are

$$1{,}000{,}000\ (1-F_k) \quad \text{letters/hr}$$

going to the residue bins, and so [15] the contribution of (*a*) to ΔS is approximately (using (5.1) and (5.2))

$$1{,}000{,}000\ (1-F_k)/(60\times 180). \quad (14.10)$$

The analysis of the contribution of (*b*) to ΔS depends upon the fact that, for our mail distribution, only destinations 1 to 114 receive 60 or more letters/hr out of 36,000 letters/hr (the input of the first loading complex of a primary subsystem). The contribution of (*b*) to ΔS is

$$0 \qquad \text{if} \quad j \leq 114. \quad (14.11)$$

If $j>114$, then the $j-114$ destinations 115, 116, . . ., j have their bins before the second loading complex receiving fewer than 60 letters/hr, leading (via (14.6)) to approximately

$$28\ (j-114)/(2+F_j-F_k) \text{ sweeps/hr.}$$

Not all of these are *extra* sweeps, however; the volume of mail involved is [16]

$$1{,}000{,}000\ (F_j-F_{114})/(2+F_j-F_k) \text{ letters/hr,}$$

which in calculating the minimum cost would receive

$$1{,}000{,}000\ (F_j-F_{114})/60(2+F_j-F_k) \text{ sweeps/hr.}$$

Thus we have a contribution of (*b*) to ΔS of approximately

$$[28(j-114)-1{,}000{,}000(F_j-F_{114})/60]/180(2+F_j-F_k) \quad (\text{if } j>114). \quad (14.12)$$

The contribution of (*c*) to ΔS arises from *two sources*. *First*, there are the type 1 destinations (if any) which receive fewer than 60 letters/hr after the second loading complex of each primary subsystem. To handle these, we define a new variable m, dependent on j and k, by

$$\text{destination } m = (\text{last destination for which } 36{,}000\ (1+F_j-F_k)f_m \geq 60). \quad (14.13)$$

The contribution to ΔS from the first source of (*c*) is then

$$0 \qquad \text{if} \quad m>j; \quad (14.14)$$

if $m \leq j$ then the $(j-m)$ destinations $m+1$, $m+2$, . . . , j are the ones under consideration, leading via (14.6) to approximately

$$28\ (j-m)/(2+F_j-F_k)$$

sweeps, of which (see the derivation of (14.12))

$$1{,}000{,}000\ (F_j-F_m)(1+F_j-F_k)/60(2+F_j-F_k)$$

are extra. Thus we have a contribution to ΔS from the first source of (*c*), of approximately

$$[28(j-m)-1{,}000{,}000(F_j-Fm)(1+F_j-F_k)/60] \div 180(2+F_j-F_k) \qquad \text{if } m \leq j. \quad (14.15)$$

Second, there are the type 2 destinations which receive fewer than 60 letters/hr after the second complex (i.e., the input from *both* complexes adds up to fewer than 60 letters/hr). To handle these, we introduce another new variable n, also dependent on j and k, by

$$\text{destination } n = (\text{last destination for which } 36{,}000\ (2+F_j-F_k)f_n > 60). \quad (14.16)$$

The contribution to ΔS from the second source of (*c*) is then

$$0 \qquad \text{if} \quad n>k; \quad (14.17)$$

reasoning as in the derivations of (14.12) and (14.15) we find the contribution to ΔS from the second source of (*c*) to be approximately

$$[28(k-n)-1{,}000{,}000(F_k-F_n)(2+F_j-F_k)/60] \div 180(2+F_j-F_k) \qquad \text{if } j \leq n \leq k \quad (14.18)$$

and to be approximately

$$[28(k-j)-1{,}000{,}000(F_k-F_j)(2+F_j-F_k)/60] \div 180(2+F_j-F_k) \qquad \text{if } n<j. \quad (14.19)$$

This completes the derivation of the approximate formula for ΔS.

At this point it is convenient to split ΔC into two parts,

$$\Delta C = \Delta_1 C + \Delta_2 C, \quad (14.20)$$

where $\Delta_1 C$ represents the effects of ΔL, ΔB, and the contribution of (*a*) to ΔS, whereas $\Delta_2 C$ represents the effects of the contributions of (*b*) and (*c*) to ΔS.

[15] It is assumed that the parameters are so chosen that all residue bins receive at least 60 letters/hr.

[16] For $1/(2+F_j-F_k)$ is the fraction of mail entering the system which enters through the *first* loading complexes of the primary subsystems.

By (8.12) and (14.10),

$$\Delta_1 C \approx 12{,}700(\Delta L) + 14(\Delta B) + 14{,}670(1{,}000{,}000(1-F_k)/(60)(180)), \qquad (14.21)$$

which, together with (14.8) and (14.9), yields

$$\Delta_1 C \approx [37.48\times 10^9 + (4.234\times 10^6)j + (3.931\times 10^6)k + (14.78\times 10^9)F_j - (52.04\times 10^9)F_k + (18.51\times 10^9)(F_k^2 - F_jF_k)]/(60)(180)\times(2+F_j-F_k), \qquad (14.22)$$

where the lower-order-of-magnitude quantity

$$(1.512\times 10^5)k(F_k-F_j)$$

has been dropped from the numerator of (14.22). No single formula can be given for $\Delta_2 C$. This is because $\Delta_2 C$ is the sum of three terms, the first of which is obtained (by multiplication by 14,670) from either (14.11) or (14.12) (according as $j \le 114$ or $j > 114$), the second of which is obtained (by multiplication by 14,670) from either (14.14) or (14.15) (according as $m > j$ or $m \le j$), from the third of which is obtained (by multiplication by 14,670) from either (14.17) or (14.18) or (14.19) (according as $n > k$ or $j \le n \le k$ or $n < j$). These diverse possibilities lead to $2\times 2\times 3 = 12$ cases; the case-by-case analysis is quite complicated, and we relegate it to appendix C, giving here only the *result:*

The cost ΔC is approximately minimized for our particular mail distribution by choosing

$$j = 120, \qquad k = 264, \qquad (14.23)$$

yielding

$$\Delta C \approx 366{,}000 \qquad (14.24)$$

so that

$$(C - C^*_{\min})/C^*_{\min} \approx 4\%. \qquad (14.25)$$

The minimum is a rather insensitive one; if j is chosen anywhere between roughly 90 and 150 then (assuming k is properly chosen) $(C-C^*_{\min})/C^*_{\min}$ will be less than 0.5 percent above its minimum.

The analysis given above (and continued in app. C) has employed the "add up" method only, deliberately avoiding any use of the "follow through" approach of section 9. We conclude this section by showing how the "follow through" method can be used (i) to reduce substantially the calculations of appendix C, and (ii) to provide rather good approximations to the optimal (j, k)-pair, (14.23). Our argument will show that j and k should be chosen to obey

$$j \ge 105, \qquad k \le 281. \qquad (14.26)$$

Use of this information would have permitted significantly less work[17] in treating cases 8 to 12 (the most difficult cases) in appendix C. Furthermore, if we regard (14.26) as suggesting $j = 105$, $k = 281$ as an approximation to an optimal choice, we find that for these values $(C-C^*_{\min})/C^*_{\min}$ is less than 0.2 percent above the value determined by (14.23).

We recall the definitions (14.1) to (14.3) of the three *types* of destinations. Our first assertion is that:

> if $10^6 f_i \le 863$, then the ith destination should not be a type 1 destination (14.27)

For, we found in the "simple" residue scheme of section 12 that if $10^6 f_i \le 863$ then mail to the ith destination should go into residue rather than receive its final sort in the primary. Since this applied after one loading complex operating at full capacity (and thus applies after the *first* complex of each of our primary subsystems here), there is even more reason for it to apply after a loading complex operating *below* capacity (such as the *second* complex in each of our primary subsystems). Thus it would be better to have the ith destination as type 3 than as type 1, and so (14.27) is proved. Of course, it might be still better to have the ith destination in type 2, and this is the next question to be considered.

If the ith destination is taken to be type 2, then (according to (9.14) and (9.17)) the cost associated with a letter to it would be denoted $C(f_i, 2)$. The formula (9.18) for $C(f_i, n_i)$ was derived assuming all primary loading complexes operating at capacity; this formula therefore provides a *lower bound* for $C(f_i, n_i)$ in our actual system, so that the special case (9.25) of (9.18) yields

$$C(f_i, 2) \ge 0.5295 + (1.337\times 10^{-3}/f_i) \text{ (if } 10^6 f_i \le 840). \qquad (14.28)$$

If, on the other hand, the ith destination is taken to be type 3, then (according to (9.11)) the appropriate cost is denoted $C(f_i)$. Except for a negligible correction due to possible violation of (9.2), the formula (9.12) still applies, yielding as in (9.24)

$$C(f_i) = 3.434 + (14/10^6 f_i) \qquad (\text{if } 10^6 f_i \ge 60). \qquad (14.29)$$

From (14.28 and (14.29) we find that

$$C(f_i, 2) - C(f_i) \ge 0 \quad \text{for} \quad 60 \le 10^6 f_i \le 455;$$

i.e., if $10^6 f_i \le 455$, then the ith destination should be taken as type 3 rather than type 2. (14.30)

From (14.30) we have $10^6 f_k > 455$, which for our particular mail distribution yields $k \le 281$.

Next we recall that in the discussion in section 9 (see (9.46)) the change from $(n_i)_{\text{opt}} = 2$ to $(n_i)_{\text{opt}} = 1$ occurred for $10^6 f_i = 1{,}614$; that is, for $10^6 f_i \ge 1{,}614$ it was better to sort a letter to the ith destination directly after it enters the system, rather than either to put it into residue or to send it on past another loader complex. In our current situation this condition must be altered to take account of the fact that the second loading complex of each subsystem operates at only a fraction $1+F_j-F_k$ of capacity.

[17] It would also yield a quick elimination of cases 1 to 5 in appendix C.

We find thus that for

$$(1+F_j-F_k)\ 10^6 f_i \geq 1{,}614 \qquad (10^6 f_i \geq 864)$$

on the basis of input to the second loading complexes only, the ith destination should be of type 1 rather than type 2.

Since $k \leq 281$ implies $F_k \leq F_{281}$, it follows that the ith destination should be of type 1 rather than type 2 if

$$(1+F_j-F_{281})\ 10^6 f_i \geq 1{,}614,$$

so that in particular we should have

$$(1+F_j-F_{281})\ 10^6 f_{j+1} < 1{,}614,$$

which for our particular mail distribution yields $j \geq 105$. Thus (14.26) is obtained.

Appendix A: Mail Distribution by Destinations

In the following table f_i denotes the fraction of mail to the ith destination, and F_i denotes the fraction of mail to the first i destinations so that

$$F_i = f_1 + f_2 + \ldots + f_{i-1} + f_i.$$

i	f_i	F_i	i	f_i	F_i	i	f_i	F_i	i	f_i	F_i
	%	%		%	%		%	%		%	%
1	3.912	3.912	56	.375	57.659	111	.172	71.887	166	.110	79.471
2	2.967	6.879	57	.371	58.030	112	.172	72.059	167	.110	79.581
3	2.893	9.772	58	.366	58.396	113	.170	72.229	168	.108	79.689
4	2.626	12.398	59	.366	58.762	114	.170	72.399	169	.108	79.797
5	2.480	14.878	60	.353	59.115	115	.166	72.565	170	.106	79.903
6	2.230	17.108	61	.345	59.460	116	.166	72.731	171	.106	80.009
7	2.057	19.165	62	.345	59.805	117	.166	72.897	172	.106	80.115
8	2.044	21.209	63	.343	60.148	118	.166	73.063	173	.106	80.221
9	1.754	22.963	64	.319	60.467	119	.164	73.227	174	.101	80.322
10	1.648	24.611	65	.319	60.786	120	.157	73.384	175	.101	80.423
11	1.635	26.246	66	.319	61.106	121	.154	73.538	176	.101	80.524
12	1.622	27.868	67	.317	61.422	122	.154	73.692	177	.101	80.625
13	1.504	29.372	68	.317	61.739	123	.154	73.846	178	.099	80.724
14	1.467	30.839	69	.315	62.054	124	.152	73.998	179	.097	80.821
15	1.379	32.218	70	.312	62.366	125	.152	74.150	180	.097	80.918
16	1.256	33.474	71	.308	62.674	126	.152	74.302	181	.095	81.013
17	1.163	34.637	72	.302	62.976	127	.150	74.452	182	.095	81.108
18	1.073	35.710	73	.297	63.273	128	.150	74.602	183	.095	81.203
19	1.053	36.763	74	.289	63.562	129	.150	74.752	184	.095	81.298
20	.982	37.745	75	.287	63.849	130	.147	74.899	185	.095	81.393
21	.827	38.572	76	.278	64.127	131	.145	75.044	186	.095	81.488
22	.814	39.386	77	.273	64.400	132	.145	75.189	187	.093	81.581
23	.773	40.159	78	.271	64.671	133	.140	75.329	188	.091	81.672
24	.734	40.893	79	.267	64.938	134	.140	75.469	189	.091	81.763
25	.713	41.606	80	.265	65.203	135	.140	75.609	190	.088	81.851
26	.706	42.312	81	.265	65.468	136	.138	75.747	191	.088	81.939
27	.702	43.014	82	.260	65.728	137	.138	75.885	192	.088	82.027
28	.655	43.669	83	.260	65.988	138	.138	76.023	193	.088	82.115
29	.644	44.313	84	.254	66.242	139	.138	76.161	194	.088	82.203
30	.623	44.936	85	.252	66.494	140	.136	76.297	195	.088	82.291
31	.615	45.551	86	.246	66.740	141	.136	76.433	196	.088	82.379
32	.604	46.155	87	.246	66.986	142	.136	76.569	197	.086	82.465
33	.580	46.735	88	.239	67.225	143	.136	76.705	198	.086	82.551
34	.571	47.306	89	.233	67.458	144	.134	76.839	199	.084	82.635
35	.551	47.857	90	.233	67.691	145	.134	76.973	200	.082	82.717
36	.538	48.395	91	.224	67.915	146	.132	77.105	201	.080	82.797
37	.534	48.929	92	.224	68.139	147	.132	77.237	202	.080	82.877
38	.534	49.463	93	.220	68.359	148	.130	77.367	203	.078	82.955
39	.534	49.997	94	.218	68.577	149	.128	77.495	204	.078	83.033
40	.508	50.505	95	.214	68.791	150	.126	77.621	205	.076	83.109
41	.508	51.013	96	.211	69.002	151	.126	77.747	206	.076	83.185
42	.506	51.519	97	.211	69.213	152	.126	77.873	207	.076	83.261
43	.501	52.020	98	.209	69.422	153	.122	77.995	208	.074	83.335
44	.491	52.511	99	.202	69.624	154	.116	78.111	209	.074	83.409
45	.491	53.002	100	.200	69.824	155	.116	78.227	210	.074	83.483
46	.483	53.485	101	.198	70.022	156	.116	78.343	211	.073	83.556
47	.465	53.950	102	.198	70.220	157	.114	78.457	212	.073	83.629
48	.461	54.411	103	.196	70.416	158	.114	78.571	213	.073	83.702
49	.457	54.868	104	.196	70.612	159	.114	78.685	214	.071	83.773
50	.445	55.313	105	.192	70.804	160	.114	78.799	215	.071	83.844
51	.407	55.720	106	.187	70.991	161	.114	78.913	216	.070	83.914
52	.401	56.121	107	.187	71.178	162	.114	79.027	217	.070	83.984
53	.394	56.515	108	.185	71.363	163	.112	79.139	218	.070	84.054
54	.390	56.905	109	.177	71.540	164	.112	79.251	219	.068	84.122
55	.379	57.284	110	.175	71.715	165	.110	79.361	220	.068	84.190

In the following table f_i denotes the fraction of mail to the ith destination, and F_i denotes the fraction of mail to the first i destinations so that

$$F_i = f_1 + f_2 + \ldots + f_{i-1} + f_i.$$

i	f_i	F_i	i	f_i	F_i	i	f_i	F_i	i	f_i	F_i
	%	%		%	%		%	%		%	%
221	.068	84.258	241	.058	85.523	261	.052	86.623	281	.045	87.609
222	.068	84.326	242	.058	85.581	262	.052	86.675	282	.045	87.654
223	.068	84.394	243	.058	85.639	263	.052	86.727	283	.045	87.699
224	.067	84.461	244	.058	85.697	264	.052	86.779	284	.045	87.744
225	.067	84.528	245	.058	85.755	265	.050	86.829	285	.045	87.789
226	.067	84.595	246	.058	85.813	266	.050	86.879	286	.043	87.832
227	.065	84.660	247	.056	85.869	267	.050	86.929	287	.043	87.875
228	.065	84.725	248	.056	85.925	268	.050	86.979	288	.043	87.918
229	.065	84.790	249	.056	85.981	269	.050	87.029	289	.043	87.961
230	.065	84.855	250	.056	86.037	270	.050	87.079	290	.043	88.004
231	.065	84.920	251	.056	86.093	271	.050	87.129	291	.043	88.047
232	.065	84.985	252	.056	86.149	272	.050	87.179	292	.043	88.090
233	.062	85.047	253	.054	86.203	273	.050	87.229	293	.043	88.133
234	.062	85.109	254	.054	86.257	274	.050	87.279	294	.043	88.176
235	.062	85.171	255	.054	86.311	275	.050	87.329	295	.043	88.219
236	.060	85.231	256	.052	86.363	276	.047	87.376	296	.043	88.262
237	.060	85.291	257	.052	86.415	277	.047	87.423	297	.043	88.305
238	.058	85.349	258	.052	86.467	278	.047	87.470	298	.043	88.348
239	.058	85.407	259	.052	86.519	279	.047	87.517	299	.043	88.391
240	.058	85.465	260	.052	86.571	280	.047	87.564	300	.041	88.432

i	Average f_i	Group %	F_i	i	Average f_i	Group %	F_i
301 to 310	0.041	0.410	88.842	626 to 650	0.013	0.325	95.942
311 to 320	.038	.380	89.222	651 to 675	.012	.300	96.242
321 to 330	.036	.360	89.582	676 to 700	.011	.275	96.517
331 to 340	.034	.340	89.922	701 to 725	.011	.275	96.792
341 to 350	.032	.320	90.242	726 to 750	.010	.250	97.042
351 to 360	.030	.300	90.542	751 to 775	.009	.225	97.267
361 to 370	.030	.300	90.842	776 to 800	.008	.200	97.467
371 to 380	.028	.280	91.122	801 to 825	.008	.200	97.667
381 to 390	.027	.270	91.392	826 to 850	.007	.175	97.842
391 to 400	.026	.260	91.652	851 to 875	.007	.175	98.017
401 to 410	.025	.250	91.902	876 to 900	.007	.175	98.192
411 to 420	.024	.240	92.142	901 to 925	.006	.150	98.342
421 to 430	.023	.230	92.372	926 to 950	.005	.125	98.467
431 to 440	.022	.220	92.592	951 to 975	.005	.125	98.592
441 to 450	.021	.210	92.802	976 to 1,000	.005	.125	98.717
451 to 460	.020	.200	93.002	1,001 to 1,050	.004	.200	98.917
461 to 470	.019	.190	93.192	1,051 to 1,100	.004	.200	99.117
471 to 480	.019	.190	93.382	1,101 to 1,150	.003	.150	99.267
481 to 490	.019	.190	93.572	1,151 to 1,200	.003	.150	99.417
491 to 500	.017	.170	93.742	1,201 to 1,250	.003	.150	99.567
501 to 525	.017	.425	94.167	1,251 to 1,300	.002	.100	99.667
526 to 550	.016	.400	94.567	1,301 to 1,350	.002	.100	99.767
551 to 575	.015	.375	94.942	1,351 to 1,400	.002	.100	99.867
576 to 600	.014	.350	95.292	1,401 to 1,450	.001	.050	99.917
601 to 625	.013	.325	95.617	1,451 to 1,600	<.001	.083	100.000

Appendix B: Analysis of Multiple Input Systems

We consider a multiple input system in which

M = number of subsystems (B1)

n = number of loading complexes in each subsystem. (B2)

In order to specify the system completely, we must also specify, for $i = 1, 2, \ldots, n$

D_i = set of destinations whose letters are dropped in the row of bins after the ith loading complex in each subsystem. (B3)

(D_n will necessarily consist of *all* the destinations.) The sets $D_1, \ldots, D_n$ may overlap (for example, some destinations may be included both in D_1 and in D_2), so that it is also necessary to deal with combinations of the D_i's such as $D_1 - D_2$ (the set of destinations included in D_1 but not in D_2). If $\mathscr{D}$ is any set of destinations (for example, $\mathscr{D}$ might be D_1 or $D_1 - D_2$), we use the notation

$F(\mathscr{D})$ = fraction of the total mail which goes to destinations in $\mathscr{D}$. (B4)

We wish now to analyze the inputs to the various loading complexes, and also the input to each subsystem as a whole. Let

I_i=ratio of the actual input to the ith loading complex of each subsystem to its maximum possible input of 36,000 letters/hr. (B5)

The quantities $I_1, I_2, \ldots, I_n$ (and thus the inputs 36,000 I_1, 36,000 I_2, . . . , 36,000 I_n) can be computed one by one from the formulas

$$\left.\begin{aligned}
&I_1=1\\
&I_2=F(D_1),\\
&I_3=F(D_2-D_1)+I_2F(D_2),\\
&I_4=F(D_3-D_2-D_1)+I_2F(D_3-D_2)+I_3F(D_3),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,\\
&I_n=F(D_{n-1}-D_{n-2}-\ldots-D_1)\\
&\qquad+I_2F(D_{n-1}-D_{n-2}\ldots-D_2)\\
&\qquad+\ldots+I_{n-1}F(D_{n-1}).
\end{aligned}\right\} \quad \text{(B6)}$$

To obtain a more compact notation for these equations, we adopt the convention $I_0=0$ and define the sets of destinations

$$D^*_{ij}=D_{i-1}-D_{i-2}-\ldots-D_j \qquad (j<i-1); \quad \text{(B7)}$$

then (B6) can be rewritten as

$$I_i=\sum_{j<i} F(D^*_{ij})I_j. \quad \text{(B8)}$$

The input to each subsystem is given by

$$\text{(input/subsystem)}=36{,}000\ (I_1+I_2+\ldots+I_n). \quad \text{(B9)}$$

Appendix C: Proofs of Results Asserted in Section 14

The following material presupposes familiarity with section 14, to which frequent reference is made. We recall that ΔC had been split into two parts,

$$\Delta C=\Delta_1 C+\Delta_2 C,$$

that a formula (14.22) had been derived for $\Delta_1 C$, but that $\Delta_2 C$ could apparently be given by any one of 12 possible formulas, leading to 12 possible cases requiring analysis.

First we write down the conditions defining the 5 cases (out of these 12) which can be treated most easily:

Case 1:	$j\leq 114$,	$m>j$,	$n<j$.
Case 2:	$j>114$,	$m>j$,	$n<j$.
Case 3:	$j\leq 114$,	$m\leq j$,	$n<j$.
Case 4:	$j>114$,	$m>j$,	$n>k$.
Case 5:	$j>114$,	$m>j$,	$j\leq n\leq k$.

These cases will be proved *logically impossible*. From (14.13) and (14.16) it can be deduced that $f_m>f_n$, so that $m<n$, ruling out cases 1 and 2. The first and third conditions of case 3 would yield, using (14.16),

$$1/600\approx f_{114}\leq f_j\leq f_n\approx 1/600(2+F_j-F_k),$$

which is impossible since

$$2+F_j-F_k\geq 2+F_j-1\geq 1+F_{114}=1.724;$$

thus case 3 is ruled out. The first and second conditions common to cases 4 and 5 yield, using (14.13),

$$1/600\approx f_{114}>f_j\geq f_m\approx 1/600(1+F_j-F_k),$$

which is impossible since (using $j\leq k$ to deduce $F_j\leq F_k$) we have

$$1+F_j-F_k\leq 1+F_k-F_k=1;$$

thus cases 4 and 5 are ruled out.

Next we describe the technique to be used in handling some of the remaining seven cases. The parameters j and k will be treated as continuous (rather than integer-valued) variables, so that calculus methods can be used in searching for the minimum of ΔC. As the two *independent* variables, it is convenient to choose not j and k, but rather

$$j \qquad \text{and} \qquad u=1+F_j-F_k. \quad \text{(C1)}$$

Then k becomes a dependent variable, and from (C1) we have

$$\partial F_k/\partial j=f_j, \quad \text{(C2)}$$

so that

$$\partial k/\partial j=(\partial F_k/\partial j)/(dF_k/dk)=f_j/f_k. \quad \text{(C3)}$$

For our purposes, it is sufficiently accurate to replace (14.13) and (14.16) by

$$f_m=1/600u, \qquad f_n=1/600(1+u), \quad \text{(C4)}$$

so f_m and f_n (hence m, n, F_m, and F_n) depend only on u and not on j, yielding

$$\partial m/\partial j=\partial n/\partial j=\partial F_m/\partial j=\partial F_n/\partial j=0 \quad \text{(C5)}$$

A necessary condition for a minimum of ΔC is

$$\partial(\Delta C)/\partial j=0,$$

and if we introduce the symbol D solely as an abbreviation for the frequently-occurring quantity

$$D=f_k(1+u)\,\partial(\Delta C)/\partial j, \quad \text{(C6)}$$

then it follows that a necessary condition for a minimum of ΔC is

$$D=0 \quad \text{(C7)}$$

We will be able to eliminate a number of the remaining seven cases by showing that they are incompatible with (C7).

A useful preliminary step is obtained by rewriting (14.8) as

$$\Delta L = -28\ (1+F_j-u)+56/(1+u),$$

rewriting (14.9) as

$$\Delta B = (760-k)+28(j+k+56(u-F_j))/(1+u),$$

noting that (14.10) can be written as

$$10^6(u-F_j)/60\times 180,$$

and then using (14.22) and (C1) to (C3) to obtain

$$f_k(1+u)\partial(\Delta_1 C)/\partial j = -17.139\times 10^5 f_j f_k(1+u) - 14f_j(1+u) - 2.195\times 10^4 f_j f_k + 392(f_j+f_k). \quad \text{(C8)}$$

The next 3 cases (out of the remaining 7) to be treated are

Case 6: $j\le 114$, $m>j$, $n>k$
Case 7: $j>114$, $m\le j$, $n<j$
Case 8: $j\le 114$, $m\le j$, $n>k$

These cases are logically possible, but we will prove that they do not give rise to the minimum of ΔC.

In case 6, $\Delta C=\Delta_1 C$, so that by (C6) and (C8),

$$D = -17.139\times 10^5 f_j f_k(1+u) - 14f_j(1+u) - 2.195\times 10^4 f_j f_k + 392(f_j+f_k). \quad \text{(C9)}$$

Since (C4) and the last condition of case 6 imply

$$f_k(1+u)>1/600,$$

we can deduce from (C9), using the fact $f_j\ge f_k$, that

$$D < -17.139\times 10^5 f_j f_k(1+u)+784f_j < (-17.139\times 10^5/600)f_j+784f_j<0,$$

so that in case 6, $D<0$, contradicting (C7).

In case 7, ΔC is obtained by adding to $\Delta_1 C$ the sum (multiplied by 14,670) of (14.12), (14.15), and (14.19). Using (C1) to (C6), we obtain

$$D = -30.722\times 10^5 f_j f_k(1+u) - 14f_j u - .22\times 10^5 f_j f_k + 2{,}660f_j + 2{,}674f_k.$$

Since the third condition of case 7, together with (C4), implies

$$f_j(1+u)<1/600,$$

while the first condition (together with $j\le k$) implies

$$f_k\le f_j\le f_{115}<1/110,$$

we have

$$D>(-30.722\times 10^5/600)f_k - 22{,}000 f_j f_k + 2{,}646f_j + 2{,}674f_k = 2{,}646f_j - 2{,}446f_k - 22{,}000 f_j f_k > 200f_j - 22{,}000 f_j f_k = 200f_j(1-110f_k)>0,$$

so that in case 7, $D>0$, contradicting (C7).

In case 8, ΔC is obtained by adding to $\Delta_1 C$ the product of (14.15) by 14,670. Using (C1) to (C6), we obtain

$$D = -30.722\times 10^5 f_j f_k(1+u) + 13.363\times 10^5 f_j f_k + 378f_j + 2{,}674f_k - 14f_j u.$$

The third condition of case 8 (together with (C4)) implies

$$f_k(1+u)>1/600,$$

so that (using the fact that $f_k\le f_j$)

$$D < (-30.722\times 10^5/600)f_j + 13.363\times 10^5 f_j f_k + 3{,}052f_j = f_j(13.363\times 10^5 f_k - 2{,}068).$$

Thus $D<0$ if $13.363\times 10^5 f_k - 2.068 < 0$: i.e., if $f_k<.00155$, or equivalently, if $k>120$. Hence, in the remainder of the discussion of case 8, we can ssume that $k\le 120$, so that $F_k\le F_{120}$. The second condition of case 8, together with (C4), yields.

$$1/600>f_j(1+F_j-F_k),$$

so that, using the fact $F_k\le F_{120}$, we obtain

$$1/600>f_j(.2662+F_j). \quad \text{(C10)}$$

By the first condition of case 8, $f_j\ge f_{114}$, so that by (C10),

$$1/600>f_{114}(.2262+F_j),$$

or $F_j\le .7142$, implying $j\le 108$,

so by (C10) $1/600>f_{108}(.2262+F_j)$

or $F_j\le .6347$ implying $j\le 73$,

so by (C10) $1/600>f_{73}(.2262+F_j)$

or $F_j\le .2949$ implying $j\le 13$,

so by (C10) $1/600>f_{13}(.2262+F_j)$,

or

$$F_j\le -0.1552,$$

which is impossible since $F_j\ge 0$. Thus case 8 is eliminated.

There are now only four cases left. We now examine

Case 9: $j\le 114$, $m\le j$, $j\le n\le k$.

It turns out that $D=0$ can occur in this case, but only in the subcase defined by

$$99\le j\le 114,\quad 187\le k\le 281. \quad \text{(C11)}$$

To prove this, we note first that in case 9, ΔC is obtained by adding to $\Delta_1 C$ the sum (multiplied by

14,670) of (14.15) and (14.18); using (C1) to (C6), this yields

$$D=-44.305\times10^5 f_j f_k u-14 f_j u-30.942\times10^5 f_j f_k +2{,}660\, f_j+2{,}674\, f_k.$$

By the second condition of case 9 (together with (C4)),

$$f_j u<1/600, \qquad \text{(C12)}$$

which together with $u\leq 1$ yields

$$D>(-44.305\times10^5/600)f_k-30.942\times10^5 f_j f_k+2{,}646 f_j +2{,}674 f_k=2{,}646 f_j-f_k(30.942\times10^5 f_j+4{,}710),$$

so that $D>0$ if

$$f_k\leq 2{,}646 f_j/(4{,}710+30.942\times10^5 f_j)$$

or equivalently, if

$$f_k\leq 1/(1{,}169.4+1.780/f_j).$$

Thus in the remainder of the discussion of case 9, we can assume that

$$f_k>1/(1{,}169.4+1.780/f_j).$$

From the first condition of case 9 we have $f_j\geq f_{114}$, and combining this with the last inequality yields

$$f_k>.00045,$$

so that $k\leq 281$ as asserted in (C11). Thus,

$$F_k\leq F_{281},$$

which together with the version

$$1/600>f_j(1+F_j-F_k)$$

of (C12) yields $1/600>f_j(1+F_j-F_{281})$, or

$$1/600>f_j(.12391+F_j). \qquad \text{(C13)}$$

Since $j\geq 1$, we have $F_j\geq F_1$ so that by (C13)

$$1/600>f_j(.12391+F_1),$$

or

$$f_j<.0102 \qquad \text{implying} \qquad j\geq 20,$$

so by (C13)

$$1/600>f_j(.12391+F_{20}),$$

or

$$f_j<.00332 \qquad \text{implying} \qquad j\geq 64,$$

so by (C13)

$$1/600>f_j(.12391+F_{64}),$$

or

$$f_j<.00229 \qquad \text{implying} \qquad j\geq 91$$

so by (C13)

$$1/600>f_j(.12391+F_{91})$$

or

$$f_j<.00208$$

implying $j\geq 99$. as asserted in (C11).

To prove the remaining part of (C11), we note that the third condition of case 9, together with (C4), yields

$$f_k(2+F_j-F_k)\leq 1/600,$$

and since $j\geq 99$ and $k\leq 281$, we have

$$f_k(2+F_{99}-F_{281})\leq 1/600,$$

yielding

$$f_k\leq .000916$$

so that $k\geq 187$ as asserted in (C11).

The 3 cases not treated so far are

Case 10: $j\leq 114$, $m>j$, $j\leq n\leq k$,

Case 11: $j>114$, $m\leq j$, $n>k$,

Case 12: $j>114$, $m\leq j$, $j\leq n\leq k$.

In all these cases, $D=0$ can occur, and no reduction like that in case 9 appeared possible. We therefore resorted to numerical exploration for these three cases and also for the range (C11) of case 9. A number of values of j were used; for each the value of k such that $D=0$ was found and the corresponding value of ΔC was computed. The results of the calculations are given in the following table, which supplies the conclusions given in (14.23) and (14.24).

*Calculations * for (C11) and cases 10 to 12*

j	k	u	ΔC	j	k	u	ΔC
20	246	0.519	466,000	114	272	.852	367,000
30	253	.587	440,000	120	264	.866	366,000
40	255	.642	421,000	125	256	.878	367,000
50	264	.685	407,000				
60	265	.723	398,000	130	256	.885	367,000
				135	253	.893	369,000
70	265	.755	388,000	140	252	.901	371,000
80	275	.778	381,000	150	245	.919	376,000
90	275	.804	375,000	160	236	.938	388,000
95	275	.815	372,000				
100	275	.825	371,000	170	226	.952	395,000
				180	213	.972	410,000
105	275	.835	370,000	194	200	.995	433,000
110	275	.847	368,000	197	197	1.000	440,000

* The quantity u can be interpreted as the ratio of the *actual* hourly input of the second loading complex of each subsystem to its *maximum* input (36,000 letters/hr).

WASHINGTON, D.C.

(Paper 63B2-11)

Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 62B, No. 2, October–December 1959

New Method of Solution for Unretarded Satellite Orbits[1]

John P. Vinti

(June 16, 1959)

An axially symmetric solution of Laplace's equation in oblate spheroidal coordinates is found, which may be used as the gravitational potential about an oblate planet. This potential, which makes the Hamilton-Jacobi equation for a satellite orbit separable, has an expansion in zonal harmonics in which the amplitudes of the zeroth and second harmonics can be adjusted to agree exactly with the values for any axially symmetric planet and a fourth harmonic which then agrees approximately with the latest value for that of the earth. The net result is therefore a reduction of the problem of satellite motion to quadratures, with use of a potential field that is much closer to the empirically accepted one for the earth than any heretofore used as the starting point of a calculation. It may thus be possible to do the gravitational theory of a satellite orbit very accurately without use of perturbation theory. The method can take into account a first harmonic in the potential, in case observations are reduced to a center which does not coincide with the center of mass of the planet.

1. Introduction

The purpose of this paper is to develop a method of solution of the central problem of satellite astronomy, viz, the theory of the effect of the oblateness of a planet on the orbit of any satellite, bound or unbound, but near enough to the planet so that the forces of other bodies may be neglected. (A preliminary note has appeared in [1].[2])

To do so, let us consider the gravitational potential V outside an axially symmetric oblate planet of mass M. We call the axis of symmetry the polar axis; this will ordinarily be indistinguishable from the axis of rotation, except for entirely negligible periodic effects. We also choose some point O on the polar axis as the "center" or origin of a coordinate system and let $\mathbf{r}$ be the corresponding position vector of a field point. With $r \equiv |\mathbf{r}|$, the declination θ as the complement of the angle between $\mathbf{r}$ and the polar axis, G the gravitational constant, M the mass of the earth, and $\mu \equiv GM$, the potential $V(r, \theta)$ can be expressed by means of an expansion in spherical harmonics:

$$V = -\mu r^{-1}\left[1 - \sum_{n=1}^{\infty} J_n (R/r)^n P_n(\sin\theta)\right]. \tag{1}$$

Here R is the equatorial radius, i.e., the radius of a section through O perpendicular to the polar axis. The notation J_n for the coefficients was introduced by Merson and King-Hele [2]. For points on the polar axis ($\theta = \pi/2$) at the distance R from the center, the ratio of any two terms in (1) is equal to the ratio of the corresponding J's.

If the planet has symmetry with respect to the equatorial plane, all the odd harmonics drop out; i.e., $J_1 = J_3 = J_5 = \ldots = 0$. (In any case, even if no such symmetry exists, the first harmonic drops out, i.e., J_1 vanishes, if O is chosen at the center of mass.) The coefficients J_2 and J_4 are given by

$$J_2 = 2J/3 = (I_p - I_t)/MR^2, \tag{2.1}$$

$$J_4 = -8D/35, \tag{2.2}$$

where the quantities J and D were introduced by Jeffreys [3], and where I_p and I_t are the

[1] This work was supported by the U.S. Air Force, through the Office of Scientific Research of the Air Research and Development Command.
[2] Figures in brackets indicate the literature references at the end of this paper.

moments of inertia of the planet about polar and transverse axes through O. For the earth, with O taken at the center of mass, Jeffreys [3] gives the values

$$J=(1637.5)10^{-6} \qquad D=(10.7)10^{-6}, \tag{3.1}$$

while King-Hele and Merson [4] derive from recent satellite observations the values

$$J=(1624.6\pm0.3)10^{-6}, \qquad D=(6\pm1)10^{-6}, \qquad J_6=(0.1\pm1.5)10^{-6}. \tag{3.2}$$

It should be noted that until very recently it has always been assumed that the earth has symmetry with respect to its equatorial plane—strictly speaking, that the geoid, or smoothed earth, has such symmetry. That is, all the odd J's have been assumed to vanish. Recently, however, there has appeared some evidence [5] for an observable asymmetry with respect to the equatorial plane. Such evidence, if confirmed, would have the following results. If O is taken at the center of mass C, then J_1 would still vanish, but $J_3, J_5, \ldots$ would have small nonvanishing values. If, on the other hand, O is taken, either by error or by design, as noncoincident with C, but still on the polar axis, then we should have $J_1 \neq 0$, with $J_3, J_5, \ldots$ still small but nonvanishing. The only conceivable difficulty that might arise with this latter choice for the center has to do with the inertial forces acting on the satellite. This difficulty is discussed and removed in the next section. A reason for deliberately choosing O as noncoincident with C might be to represent any equatorial asymmetry largely by means of a first harmonic, and such a procedure might work if the third and higher odd harmonics then turn out to be negligible in comparison with the first.

In the present paper we first derive a solution V_L of Laplace's equation, in oblate spheroidal coordinates, having axial symmetry and leading to exact separability of the Hamilton-Jacobi equation for the motion of a satellite. It contains three adjustable constants, b_0, b_1, and c, which may be so chosen that the expansion of V_L in spherical harmonics is

$$V_L=-\mu r^{-1}[1-J_2(R/r)^2P_2(\sin\theta)+J_2^2(R/r)^4P_4(\sin\theta)-J_2^3(R/r)^6P_6(\sin\theta)+\ldots$$
$$-J_1(R/r)P_1(\sin\theta)+J_1J_2(R/r)^3P_3(\sin\theta)-J_1J_2^2(R/r)^5P_5(\sin\theta)+\ldots], \tag{4}$$

where J_2 is given by (2.1) and where J_1 is the same as the J_1 in (1), viz,

$$J_1=-\delta/R. \tag{4.1}$$

δ being the northerly displacement of the center of mass C from O.

Comparison of (4) with (1) shows that the fit is exact for the zeroth harmonic, $-\mu r^{-1}$, and for the first and second harmonics. It would be exact through all harmonics if the planetary distribution of mass happened to be such that

$$J_4=-J_2^2, \quad J_6=J_2^3, \quad \ldots, \quad J_{2n}=(-1)^{n+1}J_n^2, \tag{5.1}$$

$$J_3=-J_1J_2, \quad J_5=J_1J_2^2, \quad \ldots, \quad J_{2n+1}=(-1)^nJ_1J_2^n. \tag{5.2}$$

To obtain some idea of the adequacy of the potential (4) let us consider only the even harmonics. From (3) and (4) we may construct the following table of their amplitudes, with signs.

TABLE 1

Source	Harmonic			
	(Zeroth) 10^6	-10^6J_2	-10^6J_4	-10^6J_6
Jeffreys [3]	10^6	−1090	2.45	
From fit of b_0 and c to Jeffreys' J	10^6	−1090	1.19	$-(1.30)10^{-3}$
King-Hele and Merson [4]	10^6	−1080	1.37 ± 0.23	-0.1 ± 1.5
From fit of b_0 and c to their J	10^6	−1080	1.17	$-(1.26)10^{-3}$

The fit of the zeroth and second harmonics is necessarily exact. If we fit to Jeffreys' J, the resulting fourth harmonic has an amplitude amounting to 49 percent of that given by Jeffreys; the resulting error of (4) then amounts to 1.26 ppm, corresponding to an error of 1.26 parts in 1100 for the oblateness potential. If we fit to King-Hele and Merson's J, the resulting fourth harmonic has an amplitude amounting to (86 ± 16) percent of their value for the fourth harmonic, so that the agreement is within the observational error of their determination. Even the sixth harmonic as fitted also lies within their estimates, which are, however, very hazy. In any event the resulting error of (4) then amounts to only 0.20 ppm, corresponding to an error of only one part in 5500 for the oblateness correction.

The potential function that we find in this paper, leading to Hamilton-Jacobi separability, is thus much closer to the actual potential, at every point of space, than any heretofore used as the starting point of a calculation. If the results of King-Hele and Merson [4] are confirmed, it may be possible to do orbit theory for unretarded satellites very accurately without perturbations of the solution of the Hamilton-Jacobi equation. If not, there is still a good chance of minimizing the use of perturbation theory, by distributing over the zeroth and second harmonics the secular effect of the fraction of the fourth harmonic that is not accounted for.[3]

2. Reference System

As a reference system we first choose a right-handed system of rectangular coordinates, with origin at the center O, with Oz pointing towards the planet's north celestial pole and with Ox pointing towards the planet's vernal equinox. The directions of the axes are then fixed in an inertial system. For the computation of satellite orbits a slow precession of the axis of symmetry about the celestial polar axis may be disregarded, if the precession is sufficiently slow and if the angle of the cone of precession amounts only to about $0.1''$, as in the case of the earth [3, p. 202]. Then Oz may be considered coincident with the axis of symmetry.

The forces per unit mass acting on a satellite are then, relative to such a system and in the absence of drag:

$$\mathbf{F}_1 = \text{gravitational field of the planet} = -\nabla V, \tag{6.1}$$

$\mathbf{F}_2$ = vector sum of the gravitational fields of the sun and of any moons that the planet may have, plus the inertial forces on the satellite produced by acceleration of the center of mass of the planet by the sun and the moons, (6.2)

$\mathbf{F}_3$ = inertial force on the satellite produced by acceleration of the point O in the planet, relative to its center of mass C, $= -\ddot{\boldsymbol{\delta}}$, where $\boldsymbol{\delta}$ is the vector from C to O. (6.3)

If in (1) we disregard all higher harmonics except the second and calculate $-\nabla V$, we find that the ratio of the amplitude of the oblateness force $\mathbf{F}_0$ to the total gravitational force is of the order $4.5J_2 \approx 1/200$. Thus, for the earth

$$|\mathbf{F}_0| < 5 \text{ cm/sec}^{-2}. \tag{7.1}$$

Furthermore [6],

$$|\mathbf{F}_2| \approx 10^{-4} \text{ cm/sec}^{-2} \tag{7.2}$$

and

$$|\ddot{\boldsymbol{\delta}}| = (2\pi/T_p)^2 \delta \sin \alpha \approx (1.4)10^{-20}\delta \text{ sec}^{-2} \tag{7.3}$$

for a period T_p of precession ≈ 433 days [3, p. 202] and for an angle α of the cone of precession $\approx 0.1''$. Thus

$$F_3/F_2 \approx (1.4)10^{-11}\delta \qquad (\delta \text{ in kilometers}), \tag{7.4}$$

so that F_3 is utterly negligible compared to F_2, for any conceivable displacement δ of O from C. In turn, F_2 amounts to less than 1/50,000 of the maximum oblateness force.

[3] The author is indebted to Dr. Boris Garfinkel for this suggestion.

We are thus left with $\mathbf{F}_1$ as the only appreciable force in the absence of drag, and we may choose our coordinate origin either at the center of mass C or at any near point O on the polar axis that leaves J_2 essentially unchanged but simplifies the odd harmonic part of (1).

3. Oblate Spheroidal Coordinates

We next introduce the oblate spheroidal coordinates ξ, η, and ϕ, defined by

$$x=c[(\xi^2+1)(1-\eta^2)]^{\frac{1}{2}}\cos\phi \qquad (0<\xi<\infty), \tag{8.1}$$

$$y=c[(\xi^2+1)(1-\eta^2)]^{\frac{1}{2}}\sin\phi \qquad (-1\leqq\eta\leqq 1), \tag{8.2}$$

$$z=c\xi\eta, \tag{8.3}$$

where c is a distance to be fixed later. (For many of the following calculations, the book by Morse and Feshbach [7] is a very useful reference.) The surfaces $\xi=$constant are oblate spheroids; sections perpendicular to the z-axis are circles and meridian sections are confocal ellipses of semiaxes $c(\xi^2+1)^{\frac{1}{2}}$ and $c\xi$, the focal separation being $2c$. The locus of the foci is thus a circle of radius c in the plane $z=0$. The surfaces $\eta=$constant are one-sheet hyperboloids of revolution, asymptotic to cones with vertices at the origin, and the surfaces with *right ascension* $\phi=$constant are planes containing the polar axis. Note that $\eta=\pm 1$ along Oz or $-Oz$ respectively and that for sufficiently large values of $r=(x^2+y^2+z^2)^{\frac{1}{2}}$ the coordinates ξ and η behave like $\xi\approx r/c$ and $\eta\approx\sin\theta$, where θ is the declination [7, p. 1292]. The plane $z=0$ consists of the regions $\eta=0$ and $\xi=0$, outside and inside the focal circle, respectively.

The element of arc ds is given by

$$ds^2=h_1^2d\xi^2+h_2^2d\eta^2+h_3^2d\phi^2, \tag{9}$$

where [7, p. 1292]

$$h_1=c(\xi^2+\eta^2)^{\frac{1}{2}}(\xi^2+1)^{-\frac{1}{2}} \tag{10.1}$$

$$h_2=c(\xi^2+\eta^2)^{\frac{1}{2}}(1-\eta^2)^{-\frac{1}{2}} \tag{10.2}$$

$$h_3=c(\xi^2+1)^{\frac{1}{2}}(1-\eta^2)^{\frac{1}{2}}. \tag{10.3}$$

4. Hamilton-Jacobi Equation

Henceforth all quantities proportional to the mass of the satellite are taken per unit mass of the latter. By (9) the kinetic energy of the satellite is then given by

$$T=\frac{1}{2}(h_1^2\dot{\xi}^2+h_2^2\dot{\eta}^2+h_3^2\dot{\phi}^2), \tag{11}$$

so that

$$T=\frac{1}{2}(h_1^{-2}p_\xi^2+h_2^{-2}p_\eta^2+h_3^{-2}p_\phi^2), \tag{12}$$

where the generalized momenta are

$$p_\xi=h_1^2\dot{\xi}=\partial S/\partial\xi \tag{13.1}$$

$$p_\eta=h_2^2\dot{\eta}=\partial S/\partial\eta \tag{13.2}$$

$$p_\phi=h_3^2\dot{\phi}=\partial S/\partial\phi, \tag{13.3}$$

$S(\xi, \eta, \phi)$ being the action function.

For the case of axial symmetry $V=V(\xi, \eta)$ and the Hamiltonian H becomes

$$H=(2c^2)^{-1}\left[\frac{\xi^2+1}{\xi^2+\eta^2}\,p_\xi^2+\frac{1-\eta^2}{\xi^2+\eta^2}\,p_\eta^2+\frac{p_\varphi^2}{(\xi^2+1)(1-\eta^2)}\right]+V(\xi,\eta). \tag{14}$$

The Hamilton-Jacobi equation is then

$$(\xi^2+1)(\partial S/\partial\xi)^2+(1-\eta^2)(\partial S/\partial\eta)^2+\frac{\xi^2+\eta^2}{(\xi^2+1)(1-\eta^2)}\,(\partial S/\partial\phi)^2+2c^2(\xi^2+\eta^2)V(\xi,\eta)=2c^2(\xi^2+\eta^2)\alpha_1, \tag{15}$$

where α_1 is the constant energy.

To try to separate variables in (15), we express S as

$$S=S_1(\xi)+S_2(\eta)+S_3(\phi), \tag{16}$$

whereupon (15) becomes

$$(\xi^2+1)S_1'^2+(1-\eta^2)S_2'^2+\left(\frac{1}{1-\eta^2}-\frac{1}{\xi^2+1}\right)S_3'^2+2c^2(\xi^2+\eta^2)V(\xi,\eta)=2c^2(\xi^2+\eta^2)\alpha_1. \tag{17}$$

Here

$$S_1'=dS_1/d\xi, \qquad S_2'=dS_2/d\eta \qquad S_3'=dS_3/d\phi. \tag{18}$$

From (17) it follows that $S_3'^2$ is expressible as a function of ξ and η alone, so that

$$S_3'=\alpha_3=\text{constant}. \tag{19}$$

Further inspection of (17) shows that the variables ξ and η can be separated if and only if the potential $V(\xi, \eta)$ has the form

$$V(\xi, \eta)=(\xi^2+\eta^2)^{-1}[f(\xi)+g(\eta)]. \tag{20}$$

The question then arises: Is it possible to find a solution of Laplace's equation, of the form (20), that will adequately represent the known facts about the earth's potential? As indicated in the introduction, the answer is essentially yes. We accordingly devote the next section to finding all the possible solutions of Laplace's equation that have this form.

5. Solutions of Laplace's Equation That Lead to Hamilton-Jacobi Separability

For axial symmetry, Laplace's equation in oblate spheroidal coordinates becomes [7, p. 1292]

$$\frac{\partial}{\partial\xi}\left[(\xi^2+1)\frac{\partial V}{\partial\xi}\right]+\frac{\partial}{\partial\eta}\left[(1-\eta^2)\frac{\partial V}{\partial\eta}\right]=0. \tag{21}$$

Insertion of (20) into (21) then shows that the functions $f(\xi)$ and $g(\eta)$ must satisfy the equation

$$\xi^2(\xi^2+1)f''(\xi)-2\xi(\xi^2+2)f'(\xi)+2(\xi^2+2)f(\xi)+\eta^2(1-\eta^2)g''(\eta)+2\eta(\eta^2-2)g'(\eta)-2(\eta^2-2)g(\eta)+M(\xi,\eta)=0, \tag{22}$$

where

$$M(\xi, \eta)\equiv\eta^2F(\xi)-\xi^2G(\eta) \tag{23}$$

$$F(\xi)\equiv(\xi^2+1)f''(\xi)+2\xi f'(\xi)-2f(\xi) \tag{24.1}$$

$$G(\eta)\equiv(\eta^2-1)g''(\eta)+2\eta g'(\eta)-2g(\eta). \tag{24.2}$$

By (22) a solution of (21) of the form (20) can exist only if $M(\xi, \eta)$ is separable. Now the neces-

sary and sufficient conditions that $M(\xi, \eta)$ be separable are that $F(\xi)$ and $G(\eta)$ have the forms

$$F(\xi)=4B+4C\xi^2 \tag{25.1}$$

$$G(\eta)=-4A+4C\eta^2, \tag{25.2}$$

where A, B, and C are constants and where the signs and the factor 4 are chosen with a view to later convenience. The conditions are obviously sufficient; to show that they are also necessary, note that if $M(\xi, \eta)$ is separable, there must exist functions $\phi(\xi)$ and $\psi(\eta)$ such that

$$\eta^2F(\xi)-\xi^2G(\eta)=\psi(\eta)-\phi(\xi). \tag{26}$$

On first placing $\eta=0$ in (26) and then $\xi=0$, we find respectively

$$\phi(\xi)=\psi(0)+G(0)\xi^2 \tag{27.1}$$

$$\psi(\eta)=\phi(0)+F(0)\eta^2. \tag{27.2}$$

Then, on inserting the expressions (27) into (26), we find

$$\eta^2[F(\xi)-F(0)]-\xi^2[G(\eta)-G(0)]=\phi(0)-\psi(0). \tag{28}$$

Placing either $\xi=0$ or $\eta=0$ in (28) now shows that $\phi(0)=\psi(0)$,
so that

$$\xi^{-2}[F(\xi)-F(0)]=\xi^{-2}[G(\eta)-G(0)]=4C, \tag{29}$$

where C is a constant, by a familiar argument. Thus

$$F(\xi)=F(0)+4C\xi^2 \tag{30.1}$$

$$G(\eta)=G(0)+4C\eta^2, \tag{30.2}$$

which have the general forms indicated in (25), so that the latter are also necessary.

Comparison of (24.1) and (25.1) and of (24.2) and (25.2) shows that the functions $f(\xi)$ and $g(\eta)$ must satisfy

$$(\xi^2+1)f''+2\xi f'-2f-4C\xi^2-4B=0 \tag{31.1}$$

$$(\eta^2-1)g''+2\eta g'-2g-4C\eta^2+4A=0. \tag{31.2}$$

By (23) and (25) it now follows that

$$M(\xi, \eta)=4A\xi^2+4B\eta^2. \tag{32}$$

Insertion of (32) into (22) then shows that the functions $f(\xi)$ and $g(\eta)$ must also satisfy

$$\xi^2(\xi^2+1)f''(\xi)-2\xi(\xi^2+2)f'(\xi)+2(\xi^2+2)f(\xi)+4A\xi^2= \\ \eta^2(\eta^2-1)g''(\eta)-2\eta(\eta^2-2)g'(\eta)+2(\eta^2-2)g(\eta)-4B\eta^2. \tag{33}$$

The usual argument about separability then shows that each side of (33) is equal to a constant $4K$. Thus $f(\xi)$ and $g(\eta)$ must also satisfy

$$\xi^2(\xi^2+1)f''-2\xi(\xi^2+2)f'+2(\xi^2+2)f+4A\xi^2-4K=0 \tag{34.1}$$

$$\eta^2(\eta^2-1)g''-2\eta(\eta^2-2)g'+2(\eta^2-2)g-4B\eta^2-4K=0. \tag{34.2}$$

We may avoid solving second-order differential equations, as follows: To eliminate f'', multiply (31.1) by ξ^2 and subtract the resulting equation from (34.1). To eliminate g'', multiply (31.2) by η^2 and subtract the resulting equation from (34.2). The resulting first-order equations are the linear ones

$$\xi f'-f=(\xi^2+1)^{-1}[C\xi^4+(A+B)\xi^2-K] \tag{35.1}$$

$$\eta g'-g=(\eta^2-1)^{-1}[C\eta^4-(A+B)\eta^2-K]. \tag{35.2}$$

Their most general solutions are

$$f(\xi)=K+b_0\xi+C\xi^2+(A+B-C+K)\xi \tan^{-1}\xi \tag{36.1}$$

$$g(\eta)=-K-b_1\eta+C\eta^2+\frac{1}{2}(A+B-C+K)\eta \ln\frac{1+\eta}{1-\eta} \tag{36.2}$$

where b_0 and b_1 are constants of integration. Insertion of (36.1) into (31.1) now shows that

$$B=A. \tag{37}$$

Then, with $B=A$, insertion of (36.2) into (31.2) simply yields a check.

Thus

$$f(\xi)=K+b_0\xi+C\xi^2+(2A-C+K)\xi \tan^{-1}\xi \tag{38.1}$$

$$g(\eta)=-K-b_1\eta+C\eta^2+\frac{1}{2}(2A-C+K)\eta \ln\frac{1+\eta}{1-\eta}. \tag{38.2}$$

It is easy to verify that these solutions for $f(\xi)$ and $g(\eta)$ satisfy both sets of second-order equations, viz, (31) and (34), as they should. Moreover, since $f(\xi)$ and $g(\eta)$ *must* satisfy (35) and since the expressions (36) are the most general solutions of (35), it follows that there are no other solutions for $f(\xi)$ and $g(\eta)$.

From (38) it now follows that all axially symmetric solutions of Laplace's equation that have the form (20) in oblate spheroidal coordinates are expressible as

$$V(\xi,\eta)=C+(\xi^2+\eta^2)^{-1}\left[b_0\xi-b_1\eta+b_2\left(2\xi \tan^{-1}\xi+\eta \ln\frac{1+\eta}{1-\eta}\right)\right], \tag{39}$$

where we have replaced $\frac{1}{2}(2A-C+K)$ by b_2. The Hamilton-Jacobi equation is separable, when the potential is a solution of Laplace's equation, if and only if the potential has this form in oblate spheroidal coordinates.

If as usual we take V to be zero at infinity, the constant C vanishes. We are thus left with three explicit adjustable constants, b_0, b_1, and b_2. It is at once clear, however, that b_2 must vanish if the solution is to represent the potential outside a planet, for the logarithmic term has a singularity everywhere on the polar axis, for which $\eta=\pm 1$.

We thus finally arrive at a solution of Laplace's equation which makes the Hamilton-Jacobi equation separable and which contains two explicit adjustable constants, b_0 and b_1, and one implicit adjustable constant c, the latter appearing in the equations of transformation from oblate spheroidal to rectangular coordinates. This solution is

$$V=(\xi^2+\eta^2)^{-1}(b_0\xi-b_1\eta)=b_0 \operatorname{Re}(\xi+j\eta)^{-1}+b_1 \operatorname{Im}(\xi+j\eta)^{-1}. \tag{40}$$

The reader may readily verify that $(\xi+j\eta)^{-1}$ is a solution of the Laplace equation (21) and thus that (40) is a solution. (We here use j instead of i for the imaginary unit, saving i for use in later papers as orbital inclination.)

6. Expansion of the Potential in Spherical Harmonics

From (8) we obtain

$$\xi^2-\eta^2=r^2/c^2-1, \tag{41.1}$$

$$\xi\eta=sr/c, \tag{41.2}$$

where r is the distance from the origin to the satellite and where

$$s=\sin\theta, \tag{41.3}$$

θ being the declination of the satellite. Then

$$(\xi+j\eta)^2=\xi^2-\eta^2+2j\xi\eta=r^2c^{-2}-1+2jsrc^{-1} \tag{42.1}$$

$$=r^2c^{-2}(1-2hs+h^2), \tag{42.2}$$

where

$$h\equiv -jcr^{-1}. \tag{42.3}$$

Then

$$(\xi+j\eta)^{-1}=cr^{-1}(1-2hs+h^2)^{-\frac{1}{2}}. \tag{43}$$

Since we shall choose $c<<r$, it follows that $|h|<1$ and $(1-2hs+h^2)^{-\frac{1}{2}}$ is then the generating function for the Legendre polynomials $P_n(s)$. Thus

$$(\xi+j\eta)^{-1}=cr^{-1}\sum_{n=0}^{\infty}h^nP_n(s)=cr^{-1}\sum_{n=0}^{\infty}(-j)^nc^nr^{-n}P_n(s), \tag{44}$$

from which there follow the relations

$$(\xi^2+\eta^2)^{-1}\xi=\mathrm{Re}(\xi+j\eta)^{-1}=cr^{-1}[1-c^2r^{-2}P_2(\sin\theta)+c^4r^{-4}P_4(\sin\theta)-\ .\ .\ .] \tag{45.1}$$

$$(\xi^2+\eta^2)^{-1}\eta=-\mathrm{Im}(\xi+j\eta)^{-1}=cr^{-1}[cr^{-1}P_1(\sin\theta)-c^3r^{-3}P_3(\sin\theta)+c^5r^{-5}P_5(\sin\theta)-\ .\ .\ .]. \tag{45.2}$$

From (40) and (45) it then follows that

$$V=b_0cr^{-1}[1-c^2r^{-2}P_2(\sin\theta)+c^4r^{-4}P_4(\sin\theta)-\ .\ .\ .]-b_1cr^{-1}[cr^{-1}P_1(\sin\theta)-c^3r^{-3}P_3(\sin\theta)+\ .\ .\ .] \tag{46}$$

7. Evaluation of the Adjustable Constants

Comparison of (46) with (1) shows that to fit the zeroth harmonic we must choose

$$b_0c=-\mu, \tag{47.1}$$

that to fit the second harmonic we must then choose

$$c^2=R^2J_2, \tag{47.2}$$

and that to fit the first harmonic we must choose

$$b_1c^2=-\mu RJ_1=\mu\delta. \tag{47.3}$$

That is all the flexibility that (46) permits; the higher harmonics are then all fixed, as indicated in the relations (5), which follow from (46).

These choices lead to the form (4) for the potential, which has been amply discussed in the introduction. For $R=6378.388$ km we find from (47.2) and table 1:

$$c=210.7 \text{ km} \qquad \text{(Jeffreys)}, \tag{48.1}$$

$$c=209.9 \text{ km} \qquad \text{(King-Hele and Merson)}. \tag{48.2}$$

Although eqs (8) appear to place this constant c in so central a position in the theory as to make numerical predictions depend strongly on its accuracy, it is not really so. Actually we shall see that a knowledge of the value of c is needed only to find the deviations of the orbit from a conic section.

From (40) and (47) we obtain as our working expression for V, in oblate spheroidal coordinates:

$$V=-\mu c^{-1}(\xi^2+\eta^2)^{-1}(\xi+\eta\delta/c). \tag{49}$$

8. Formal Solution of the Hamilton-Jacobi Equation

Insertion of (19) and (49) into (17) now results in the fully separated equation

$$(\xi^2+1)S_1'^2-(\xi^2+1)^{-1}\alpha_3^2-2\mu c\xi-2\alpha_1 c^2\xi^2=-(1-\eta^2)S_2'^2-(1-\eta^2)^{-1}\alpha_3^2+2\mu\delta\eta+2\alpha_1 c^2\eta^2=k. \tag{50}$$

By the usual argument, each side of (50) is equal to a constant k.

At this point it is necessary to discuss complete and incomplete orbits. We define a complete orbit as the orbit that would be traversed by the satellite when the time t is allowed to run from $-\infty$ to $+\infty$, on the assumption that the equations defining the orbit hold for all values of t. In the cases of a planetary satellite or of a meteor coming in from infinity with its orbit being bent around the planet without hitting it, the actual physical orbit is complete. Any projectile fired from a gun and returning to earth, however, exemplifies an incomplete bound orbit. Any missile which escapes from the earth exemplifies an incomplete unbound orbit, as does any meteor with energy $\alpha_1>0$ which hits the earth. Note that the complete orbit may or may not intersect the planet. If it does, we define it as that complete orbit which would be described by the satellite, in the absence of planetary matter, but in a gravitational field corresponding to (49). For $r>c$, this field can be represented by the multipole field (46).

By (50) it follows that

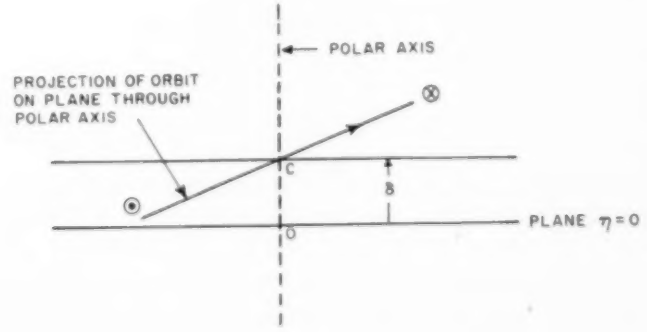
$$k+\alpha_3^2=-(1-\eta^2)S_2'^2-\eta^2(1-\eta^2)^{-1}\alpha_3^2+2\mu\delta\eta+2\alpha_1 c^2\eta^2. \tag{51}$$

Thus if the plane $\eta=0$ is reached by the complete orbit

$$k+\alpha_3^2\leqq 0. \tag{52}$$

In the case where δ is chosen to be zero, the plane $\eta=0$ is the ordinary equatorial plane through the center of mass of the planet. Thus, if this plane is reached by the complete orbit, (52) is satisfied.

It is clear that (52) is only a necessary condition that the complete orbit reach the plane $\eta=0$. It is difficult to find a corresponding sufficient condition. By analogy with the Kepler problem (r^{-1} potential), we may suspect that when $\delta=0$ the equatorial plane will always be reached by the complete orbit of any bound satellite. If $\delta\neq 0$, however, the plane $\eta=0$ may not be reached if the orbit is very close to the equator. This latter circumstance is illustrated in figure 1.



The difficulties in finding a condition sufficient for $\eta=0$ to be reached are illustrated by the existence of cases where the complete orbit may not reach either the region $\eta=0$ outside the focal disk or the region $\xi=0$ inside it. For example, if a particle is dropped from a point on the polar axis into a matter-free region satisfying (49), so that $\alpha_3=0$, $\eta_0=\pm 1$, $\dot{\eta}_0=0$, it follows (with $\delta=0$) that the complete orbit is a piece of the polar axis, penetrating neither $\xi=0$ nor $\eta=0$.

In the rest of the paper we exclude the possibility of orbits that lie only on the polar axis, thus ruling out, e.g., the case of a missile fired vertically from either pole. We then consider only those orbits whose completions either reach the plane $\eta=0$ or reach at least one genuine extremum η_1 for which $\dot{\eta}=0$, $\ddot{\eta}\neq 0$.

From (50) it now follows that

$$S_1'(\xi)=\pm(\xi^2+1)^{-1}[\alpha_3^2+(k+2\mu c\xi+2\alpha_1 c^2\xi^2)(\xi^2+1)]^{\frac{1}{2}} \tag{53.1}$$

$$S_2'(\eta)=\pm(1-\eta^2)^{-1}[-\alpha_3^2+(-k+2\mu\delta\eta+2\alpha_1 c^2\eta^2)(1-\eta^2)]^{\frac{1}{2}}, \tag{53.2}$$

where by (13), (16), and (18) the upper sign is to be used whenever the corresponding coordinate is increasing and the lower sign whenever it is decreasing. Also

$$S_3'(\phi)=\alpha_3. \tag{19}$$

From (53) and (19) it then follows that

$$S=\alpha_3\phi\pm\int_{\xi_1}^{\xi}(\xi^2+1)^{-1}[\alpha_3^2+(k+2\mu c\xi+2\alpha_1 c^2\xi^2)(\xi^2+1)]^{\frac{1}{2}}d\xi$$
$$\pm\int_{\eta_1}^{\eta}(1-\eta^2)^{-1}[-\alpha_3^2+(-k+2\mu\delta\eta+2\alpha_1 c^2\eta^2)(1-\eta^2)]^{\frac{1}{2}}d\eta+\text{const.} \tag{54}$$

Before defining the lower limits ξ_1 and η_1, it is well to note that eq (54) is somewhat deceptive in appearance, because of the presence of the "small" distance c in terms that are clearly very important. Let us now remove this difficulty, as follows. Replace the variable ξ by the variable ρ, where

$$\rho\equiv c\xi. \tag{55}$$

By (41) it then follows that

$$\rho^2=r^2-(1-\eta^2)c^2 \tag{56.1}$$

$$\rho\eta=r\sin\theta \tag{56.2}$$

$$\xi^2+1=(\rho^2+c^2)c^{-2}, \tag{56.3}$$

so that $\rho\approx r$ on the physically realizable part of any complete orbit. On substituting (55) into (54) and dropping the constant term, we obtain

$$S=\alpha_3\phi\pm\int_{\rho_1}^{\rho}(\rho^2+c^2)^{-1}[c^2\alpha_3^2+(k+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{\frac{1}{2}}d\rho$$
$$\pm\int_{\eta_1}^{\eta}(1-\eta^2)^{-1}[-\alpha_3^2+(-k+2\mu\delta\eta+2\alpha_1 c^2\eta^2)(1-\eta^2)]^{\frac{1}{2}}d\eta, \tag{57}$$

where

$$\rho_1=c\xi_1. \tag{57.1}$$

If the complete orbit intersects the plane $\eta=0$, we choose the lower limit η_1 to be zero, corresponding to $\theta=0$, in accord with the usual canonical treatment [8] of a bound orbit for the case $V=-\mu r^{-1}$; this appears to be a good choice for an unbound orbit also. If the complete orbit does not intersect the equatorial plane, we choose η_1 as an extremum of η on the complete orbit. It will be a zero of the η-integrand satisfying $\eta\geqq -1$.

To choose a value for ρ_1, we reason as follows. The ρ-integrand of (57) has the factor

$$F(\rho)=c^2\alpha_3^2+(\rho^2+c^2)f(\rho), \tag{58.1}$$

where

$$f(\rho)=2\alpha_1\rho^2+2\mu\rho+k. \tag{58.2}$$

The appropriate real zeros of $F(\rho)$ will be almost equal to those of $f(\rho)$, which always has two in any physically realizable case. If $\alpha_1\geqq 0$, one zero of $f(\rho)$ will be positive and the other negative (approaching $-\infty$ as α_1 approaches zero). If $\alpha_1<0$, both zeros of $f(\rho)$ will be positive. Any other real zeros of $F(\rho)$ will occur for inappropriate values of ρ, of the order c. Thus for unbound orbits, we choose ρ_1 to be the largest positive zero of $F(\rho)$, this being the only one accessible to complete unbound orbits. For bound orbits, we choose ρ_1 to be the next to the largest positive zero of $F(\rho)$, equal approximately to the smaller zero of $f(\rho)$. In either case ρ_1 is then the minimum value of ρ dynamically accessible to the complete orbit; the corresponding spheroid may or may not lie within the earth. (Any value of ρ less than ρ_1 as just defined is dynamically inaccessible to the complete orbit in the field $V(\xi,\eta)$, even if the planetary matter could be removed without change in $V(\xi,\eta)$: to reach such values the orbit would have to pass through regions in which $F(\rho)$ is negative and p_ξ complex.) Of course $\rho_1>R$ and thus outside the earth for any ordinary satellite in a bound orbit.

Whenever the plane $\eta=0$ is intersected by the complete orbit, we have seen that $k\leqq-\alpha_3^2$, so that we can then introduce a real α_2 such that

$$k=-\alpha_2^2, \tag{59.1}$$

where we can assume

$$\alpha_2>0 \tag{59.2}$$

without loss of generality. Then

$$\alpha_2^2\geqq\alpha_3^2. \tag{59.3}$$

If we put $\delta=0$ and $c=0$ in (57) and then use $\rho=r$, $\eta=\sin\theta$, which would then follow from (56), and also use (59.1), we find that (57) reduces [8] to the correct expression for the action function for the case $V=-\mu r^{-1}$. Thus the terms in (57) that contain c^2 or δ give rise to deviations from a conic-sectional orbit.

From (13.3) and (19) it follows that the constant α_3 in (57) is the z-component of angular momentum. The constant k has no name, but is related to the magnitude of the total angular momentum, which is, however, not conserved in the non-central field produced by oblateness. To see the connection, note that whenever $k\leqq-\alpha_3{}^2$, we can write (59.1) and (59.3). Then when δ and c^2 are both allowed to approach zero in (57), α_2 approaches the magnitude of the total angular momentum, which is then conserved. (This result follows most easily from the fact already mentioned, viz, that (57) then reduces to the correct action function for the case $V=-\mu r^{-1}$.)

9. Kinetic Equations of Motion

Implicit equations for ρ, η, and ϕ as functions of t are now given by

$$\partial S/\partial\alpha_1=t+\beta_1 \tag{60.1}$$

$$\partial S/\partial k=\beta_2' \tag{60.2}$$

$$\partial S/\partial\alpha_3=\beta_3 \tag{60.3}$$

where t is the time and the β's are constants. Since the lower limit ρ_1 in (57) is a function of α_1, k, and α_3, the derivative of the ρ-integral with respect to any of them is the sum of an integral and a term like $\mp I_\rho(\rho_1)\partial\rho_1/\partial\alpha_1$. Here $I_\rho(\rho)$ is the ρ-integrand, which vanishes for ρ_1, so that this second term vanishes. If the complete orbit intersects the plane $\eta=0$, so that η_1 can be taken to be zero, on second term arises in a derivative of the η-integral; if not, η_1 is to be

taken as an extremum value of η on the complete orbit and the second term again vanishes, since η_1 is then a zero of the η-integrand.

The kinetic equations are thus

$$\pm\int_{\rho_1}^{\rho}\rho^2[c^2\alpha_3^2+(k+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{-\frac{1}{2}}d\rho\pm c^2\int_{\eta_1}^{\eta}\eta^2[-\alpha_3^2+(-k+2\mu\delta\eta+2\alpha_1c^2\eta^2)(1-\eta^2)]^{-\frac{1}{2}}d\eta=t+\beta_1, \tag{61.1}$$

$$\pm\tfrac{1}{2}\int_{\rho_1}^{\rho}[c^2\alpha_3^2+(k+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{-\frac{1}{2}}d\rho\mp\tfrac{1}{2}\int_{\eta_1}^{\eta}[-\alpha_3^2+(-k+2\mu\delta\eta+2\alpha_1c^2\eta^2)(1-\eta^2)]^{-\frac{1}{2}}d\eta=\beta_2', \tag{61.2}$$

$$\phi\pm c^2\alpha_3\int_{\rho^1}^{\rho}(\rho^2+c^2)^{-1}[c^2\alpha_3^2+(k+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{-\frac{1}{2}}d\rho$$
$$\mp\alpha_3\int_{\eta_1}^{\eta}(1-\eta^2)^{-1}[-\alpha_3^2+(-k+2\mu\delta\eta+2\alpha_1c^2\eta^2)(1-\eta^2)]^{-\frac{1}{2}}d\eta=\beta_3. \tag{61.3}$$

If we restrict considerations to those cases where the plane $\eta=0$ is intersected by the complete orbit, then $k\leqq\alpha_3^2$. We can then substitute $k=-\alpha_2^2\leqq-\alpha_3^2$ in (57) and replace (60.2) by

$$\partial S/\partial\alpha_2=\beta_2. \tag{62}$$

From (57), (60.1), (60.3), and (62) there then follow the kinetic equations

$$\pm\int_{\rho_1}^{\rho}\rho^2[c^2\alpha_3^2+(-\alpha_2^2+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{-\frac{1}{2}}d\rho\pm c^2\int_0^{\eta}\eta^2[-\alpha_3^2+(\alpha_2^2+2\mu\delta\eta$$
$$+2\alpha_1c^2\eta^2)(1-\eta^2)]^{-\frac{1}{2}}d\eta=t+\beta_1, \tag{63.1}$$

$$\mp\alpha_2\int_{\rho_1}^{\rho}[c^2\alpha_3^2+(-\alpha_2^2+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{-\frac{1}{2}}d\rho\pm\alpha_2\int_0^{\eta}[-\alpha_3^2+(\alpha_2^2+2\mu\delta\eta+2\alpha_1c^2\eta^2)(1-\eta^2)]^{-\frac{1}{2}}d\eta=\beta_2, \tag{63.2}$$

$$\phi\pm c^2\alpha_3\int_{\rho_1}^{\rho}(\rho^2+c^2)^{-1}[c^2\alpha_3^2+(-\alpha_2^2+2\mu\rho+2\alpha_1\rho^2)(\rho^2+c^2)]^{-\frac{1}{2}}d\rho$$
$$\mp\alpha_3\int_0^{\eta}(1-\eta^2)^{-1}[-\alpha_3^2+(\alpha_2^2+2\mu\delta\eta+2\alpha_1c^2\eta^2)(1-\eta^2)]^{-\frac{1}{2}}d\eta=\beta_3. \tag{63.3}$$

Future papers will be devoted to evaluating the integrals in these equations, solving them simultaneously for ρ, η, and ϕ as functions of time, and deriving the astronomical results that they imply, both for bound orbits and unbound orbits.

10. References

[1] J. P. Vinti, Phys. Rev. Letters **3,** No. 1 (July 1959).
[2] R. H. Merson and D. G. King-Hele, Nature **182,** 640 (Sept. 1958).
[3] H. Jeffreys, The earth, 3d ed (Cambridge Univ. Press, Cambridge, England, 1952). See p. 127 at bottom; p. 129, eq (5); and p. 132, eq (13).
[4] D. G. King-Hele and R. H. Merson, Nature **183,** 881 (March 1959).
[5] J. A. O'Keefe, A. Eckels, and R. K. Squires, Science **129,** 565 (Feb. 1959).
[6] J. A. O'Keefe, Astron. J. **62,** 265 (1957).
[7] P. M. Morse and H. Feshbach, Methods of theoretical physics, vols. I and II (McGraw-Hill Book Co., Inc., New York, N.Y., 1953).
[8] W. M. Smart, Celestial mechanics, p. 143 (Longmans, Green & Co., New York, N.Y., 1953).

WASHINGTON, D.C. (Paper 63B2–12)

Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 63B, No. 2, October-December 1959

Effect of Sudden Water Release on the Reservoir Free Outflow Hydrograph[1]

Vujica M. Yevdjevich[2]

(July 9, 1959)

The free outflow hydrograph is studied in the case of a sudden water release from a reservoir. The outflow hydrograph is called free when it is not affected by the tailwater levels. Both the effect of the steep negative wave, created by sudden water release, and the effect of the flow resistance are analyzed through the use of a fictitious inflow hydrograph. The water accelerated by the steep negative wave movement, whether or not the flow resistance is taken into account, gives this fictitious hydrograph. The procedures are given for the computation of the wave celerities and the new water velocities along the reservoir, for the computation of wavefront heights and the maximum outflow discharges, and for the determination of the fictitious and total inflow hydrograph. The examples are given for the pyramidal reservoirs of rectangular and parabolic cross sections. A general procedure for the determination of the reservoir free outflow hydrograph in an approximate form is derived.

1. Introduction

1.1. Definition of the Free Outflow Hydrograph

If the outflow hydrograph from a reservoir is unaffected by the tailwater levels or if such influence is negligible, it will be called the free outflow hydrograph. In that case both the outflow hydrograph and the downstream propagation of the wave created by the sudden release of water can be analyzed separately, first the hydrograph and then the wave propagation.

The reservoir free outflow hydrograph in the case of sudden water release depends on: (*a*) The area and shape of the outlet; (*b*) the type of flow (free surface flow or flow under pressure); (*c*) the shape characteristics of water body (storage function, cross sections, etc); (*d*) the inflow hydrograph; (*e*) the time interval in which a full opening of the outlet is created; (*f*) the movement of negative waves, especially of steep negative waves; (*g*) the resistance to flow along the reservoir.

The free outflow hydrograph can be approximated by taking into consideration only the most influential of these factors, (*a*) to (*d*). Any classical reservoir routing procedure will suffice for the computation. The analytical solutions are possible in cases when the outflow rating curve, the storage function and the inflow hydrograph are given and can be approximated by simple expressions mathematically tractable [3].[3] In that case the opening time is assumed as zero and the effects of steep negative wave movement and of flow resistance are neglected, assuming that at the very moment of fast opening the water in the reservoir was instantaneously accelerated to give maximum discharge. This assumption is the source of the errors in the shape of the free hydrograph, if the steep negative wave is sufficiently high, as when the opening is large in comparison with the reservoir cross sections.

1.2. Purpose of the Study

This study seeks the development of a simple procedure for the computation of free outflow hydrograph, when the movement of the steep negative wave and the flow resistance are taken into consideration. The procedure must be suitable for use with the available data, to the accuracy demanded of the resulting hydrographs, and it must require a reasonable computation time.

1.3. Scope of the Study

This study will investigate the influence of the steep negative wave and of flow resistance on the shape of the free outflow hydrograph. The knowledge of the shape of reservoir (cross sections) and of its roughness coefficients is needed. Since the opening time is very short in comparison with the outflow time of the major bulk of water, it can be taken as nearly zero.

1.4. Basic Approach to the Problem

The shape of the free outflow hydrograph, computed by neglecting the effect of the steep negative wave and of flow resistance, is to be considered as the first approximation of the true outflow hydrograph. This first shape is sufficiently accurate for small negative wave heights in many practical cases, because the accuracy of background data, the routing procedure and the accuracy demanded do not justify more precise computation. The first approximate shape of the free outflow hydrograph is thus con-

[1] This work was prepared under a National Bureau of Standards contract with The American University, with the sponsorship of the Corps of Engineers, Army Map Service.
[2] U.S. Geological Survey, Washington 25, D.C.
[3] Figures in brackets indicate the literature references at the end of this paper.

ceived as a gradual emptying of the body of water, starting from already accelerated water [3]. If the neglected factors were taken into consideration and if more data about the reservoir were available, the subsequent approximations would be closer to the true hydrograph.

Two approaches can be used in the computation of the subsequent approximations of the true outflow hydrograph: (1) To develop a special procedure for computing the hydrograph directly from all given data and (2) to correct the previous approximate hydrograph by introducing the effects of new factors. In the second approach the effects are represented by a new fictitious inflow hydrograph. The variables are time, the starting reservoir level, the characteristics of the steep negative wave, the length of reservoir, the flow resistance, etc.

The new approximation of the true hydrograph can be computed from the first approximation by simple transformation using the fictitious function of inflow. Next, the sums of the real inflow hydrograph and fictitious inflow hydrograph can be made. The new inflow hydrographs can be applied directly for the computation of the other approximations of the true hydrograph. This procedure will be used in this study.

1.5. Effect of the Steep Negative Wave and of Flow Resistance on the Free Outflow Hydrograph

The outflow from a reservoir in the case of sudden water release does create a steep negative wave which travels upstream along the reservoir.

Every increase of outflow discharge also increases the head losses along the body of water.

When a steep negative wave is created, its movement is influenced by:

(*a*) The change of the wave area A_1 and of its shape (A_1 is the vertical projection of the wavefront) and the change of the total cross section area A or of the area under the wave ($A-A_1$), figure 1, which change the propagation celerity; (*b*) the change of the direction and the width of reservoir, of the shape and the roughness of its shores, which cause and affect the reflection and the refraction of the wave; (*c*) the resistance to the acceleration of water due to the movement of the wave, and the resistance of the accelerated water moving downstream between the wavefront and the outlet, which decrease the volume and the kinetic energy of the accelerated water. The flow resistance decreases the outflow discharges in the first part of the hydrograph and increases them in its tail part.

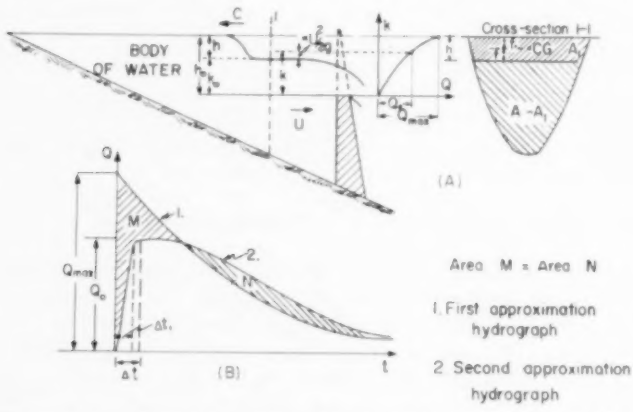


FIGURE 1. *Reservoir outflow hydrograph in case of sudden water release.*

(A) Schematic reservoir with negative wave movement; (B) two outflow hydrographs: 1—hydrograph with already accelerated water to Q_{max} at the moment of release; 2—hydrograph greatly influenced by the steep negative wave.

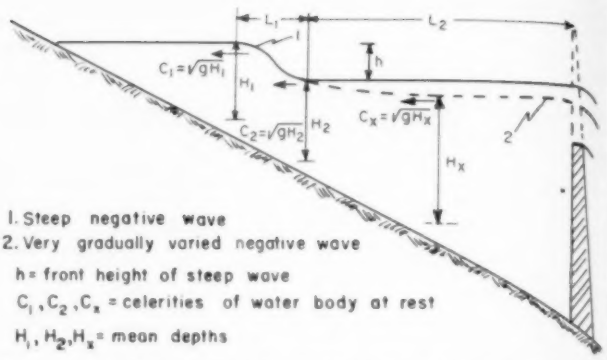


FIGURE 2. *Two types of negative waves in case of sudden water release.*

Steep negative wave and long negative wave very gradually varied.

As the propagation celerity of the negative wave gradually decreases in the upstream direction, because of the gradual decrease of mean depth in a reservoir with sloping bottom, the accelerated discharge per unit of time thus correspondingly decreases. As the level at the outlet would be higher at any previous moment than the level corresponding to the accelerated arriving discharge at the next moment, a new level adaptation must be made to satisfy the continuity equation by a lowering of the level at the outlet, which causes the additional acceleration of water in the reservoir. The volume of water involved in this lowering is taken from surface layers of the body of water between the outlet and the steep negative wave. Therefore, a continuous propagation of small negative waves, changing very gradually, takes place from the outlet upstream. For each time unit a small drop in level at the outlet is created, which propagates upstream with the celerity $C=\sqrt{gH_x}$ where H_x is the mean depth (fig. 2). As H_x near the steep negative wave is lower than the mean depth of the lowest part of steep negative wave, the small waves can not reach the steep wave. As the process is continuous, the wave movement can be conceived as a group of very small waves making a new long wave, very gradually changing (fig. 2), with H_x decreasing.

1.6. Approximations of the Free Outflow Hydrograph

The combination of the movement of two types of negative waves, of a very steep wave (depression) and of a very long wave corresponding to gradually varied unsteady flow (with quite opposite characteristics), the complex reflections and refractions

of waves in a natural body of water, and the resistance to the flow make the exact computation of the free outflow hydrograph very complicated. Therefore, the solutions must be obtained in approximations of a different degree of accuracy. The first approximation neglected the movement of negative waves and the flow resistance (as in [3]); the second approximation will include the addition of the steep wave effect; the third approximation will take account of the additional effect of flow resistance. These two approximations are treated in the subsequent part of this paper. Further approximations can be considered when practically feasible.

2. Effect of the Steep Negative Wave on the Free Outflow Hydrograph

2.1. General Assumptions

As the breach of a dam or an accidental opening of some types of large gates are relatively fast processes (a matter of few seconds), the negative wave created by sudden water release is a steep wave traveling upstream. Quite different is the case when the opening of outlets is very slow, the outflow rating curve changing with time as the opening becomes larger. Q is a function of opening dimensions and of time. This type of opening will not be studied here.

The steep negative wave undergoes three types of changes as it travels upstream: (a) Reflections and refractions; (b) decrease of propagation celerity due to the change of cross section; and (c) loss of energy due to the resistance.

The reflection and refraction of the steep negative wave are complex because of the factors which essentially change continuously along the reservoir. The process is, however, somewhat less complex in channels of nearly regular shape. The effects of the reflection and refraction of the steep negative wave will be assumed in the second and third approximations as negligible for the valley reservoirs.

The effect of the negative wave ends when its propagation celerity C in water at rest equals the mean flow velocity V already existent ($C=V$), or when the wave arrives at the shallow regions, if water is in rest during the wave movement.

The effect of resistance will be neglected in the second approximation, which is justified especially for large bodies of water and moderate heights of steep waves. It will be considered in the third approximation (sec. 3 of this paper).

The main factor in this second approximation is the additional influence of the existence and movement of steep negative wave.

The flow through reservoirs or lakes, where the mean velocities are relatively low and the cross sections change gradually, is a gradually varied flow, either steady or unsteady.

Before the analysis of the effect of the steep negative wave on the outflow hydrograph is begun, two problems will be studied: (1) The propagation of the wave along the water body and (2) the steep wave height and the outflow discharge shortly after the sudden water release.

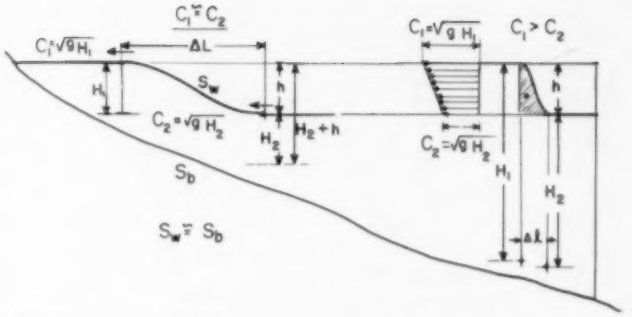


FIGURE 3. *Schematic representation of the transformation on the front shape of steep negative wave, due to different celerities of wave elements.*

The shape of a steep negative wave moving along the body of water is not stable (fig. 3). The reason is that the parts of wavefront travel with different celerities. The steepness of the wave decreases with the distance traveled, and the final shape will be obtained when the slope of the wavefront is essentially equal to the slope of the bottom. Because of these differences in celerities of different points of the wave, it is necessary always to define C (as that of the highest point, of the lowest point, of the center of gravity, of the mean height of wave, etc.). The approximations in this study deal with the schematic negative wave, considered as a discontinuous steep wave.

2.2. Mean Celerity of the Steep Negative Wave in a Cross Section

The steep negative wave arrives at a cross section of the body of water, X distance from the outlet, at time t from the opening time, with wavefront height h. The celerity of the wave in water at rest is C at X. The mean velocity of water because of inflow is V. The cross section characteristics are (fig. 1): Area A before the wave arrival; area A_1 of the wavefront height; distance f of the center of gravity CG of area A_1 from the new level (or from the starting level $f_1=h-f$); area $A-A_1$, after the wave has passed. The effect of the bottom slope is indirectly introduced by the change of A, A_1, f, f_1, and $A-A_1$.

Two equations, continuity and momentum, give the celerity of the steep negative wave and the increase of the mean water velocity in a cross section of reservoir, due to the acceleration caused by the wave movement, as functions of section characteristics [1, 2]. For a small time interval dt the celerity can be considered constant in the vicinity of the cross section at X, regardless of the bottom slope which has little influence on the cross section characteristics for short path length (fig. 4).

The steep negative wave travels upstream a distance $(C-V)dt$ during the time dt (fig. 4), and the water travels downstream the distance $(V+U)dt$. An observer, traveling at the section X with the velocity V downstream, after time dt, will be at the section $X-Vdt$, or a distance Vdt downstream. To him the distance traveled by the wave in the

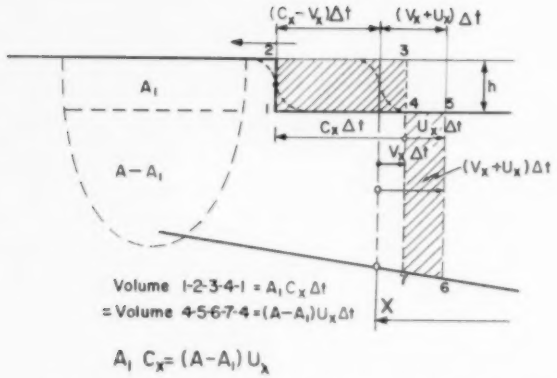


FIGURE 4. *Scheme for the derivation of continuity equation in case of movement of steep negative waves.*

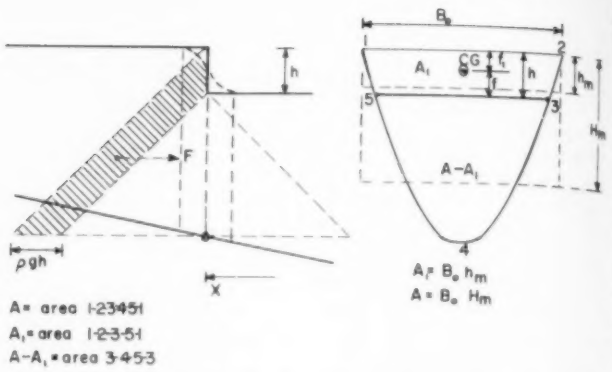


FIGURE 5. *Scheme for the derivation of momentum equation in case of steep negative wave movements.*

upstream direction is Cdt and the distance traveled by the water downstream is Udt.

The continuity equation gives for the observer (or also in case $V=0$)

$$A_1 C\,dt = (A - A_1) U dt,$$

so that generally for any X

$$C = U\left(\frac{A}{A_1} - 1\right), \tag{1}$$

where C is the propagation celerity of the steep negative wave in water at rest and $C-V$ is the propagation celerity in moving water with velocity V ($V<C$). Equation (1) can be obtained also by equating the total water discharge entering the small volume Adx, to the discharge leaving that volume

$$VA + (C - V)A_1 = (V + U)\,(A - A_1).$$

For the momentum equation, it is assumed that the acceleration of water from the velocity V to the velocity $V+U$ of the volume of water $(A-A_1)Cdt$, in time dt and at the section X (figs. 4 and 5) is made without loss of energy.

The volume element $(A-A_1)Cdt$, which is accelerated from V to the velocity $V+U$ between times t and $t+dt$, has a momentum $(A-A_1)CUdt$ and a rate of change of momentum $(A-A_1)CU$. The force F which creates this change is the difference of pressure on the areas A and $A-A_1$, (fig. 5), $F = gh(A-A_1) + gfA$, with f already defined.

This force is equal to the rate of change of momentum, so that

$$CU = gh + gf\,\frac{A_1}{A - A_1}.$$

The celerity C at any section in water at rest, after using eq (1), is

$$C = \sqrt{gh(a-1) + gf}, \tag{2}$$

where $a = A/A_1$. The ratio $a>1$, and theoretically can be between the limits 1 and infinity. If h increases, $(a-1)$ decreases faster than h and f, so that the mean celerity of the wavefront decreases with the increase of h. With the mean depths of areas A and A_1, $H_m = A/B_0$ and $h_m = A_1/B_0$ (fig. 5), with B_0 the width at the surface level, and putting $f_1 = h - f$, eq (2) becomes

$$C = \sqrt{g\left(H_m \frac{h}{h_m} - f_1\right)}. \tag{3}$$

When h/hm is nearly 1 and f_1 can be neglected in comparison with H_m (for small values of h) then

$$C = \sqrt{gH_m}, \tag{4}$$

which is the known expression for the propagation of small surface disturbances in shallow water, with the mean depth H_m.

The increase of the mean water velocity is, therefore, from eqs (1) and (2)

$$U = \frac{1}{a-1}\sqrt{gh(a-1) + gf}, \tag{5}$$

or from eqs (1) and (3)

$$U = \frac{h_m}{H_m - h_m}\sqrt{g\left(H_m \frac{h}{h_m} - f_1\right)}, \tag{6}$$

which approaches zero when h approaches zero.

The discharge $P_w = U(A - A_1)$ is the accelerated discharge by the movement of the steep negative wave, which will be considered as the fictitious inflow in the reservoir.

As H_m, h_m, and f_1 are functions of cross section characteristics and of height of wavefront h, both C and U are functions of those parameters. The cross section characteristics are functions of the water level. For a given section, level, and height h the values C and U are uniquely determined.

2.3. Computation of the Initial Front Height of Steep Negative Wave and of the Maximum Outflow Discharge

If an opening in a reservoir is created with constant area and shape, its hydraulic characteristics are defined by the outflow rating curve, generally of the type $Q=bk^r$, where b and r depend on the area, shape and position of the opening, as well as on the type of flow, with k the total energy head above the level of the outflow zero (fig. 1). The opening height, as the difference between the level at the moment of opening and the level of outflow zero, is h_0. It is the sum of two heights: h, the front height of the steep negative wave and k_0, the difference of the level just upstream from the outlet (where the depression due to the outflow is negligible) and the level of outflow zero, so that $h_0=h+k_0$. The energy head is

$$k=k_0+\frac{\alpha(V+U)^2}{2g}, \tag{7}$$

where V is the velocity near the opening before it is created and α is the kinetic energy correction factor.

For the small time interval, Δt, after the opening, the water flowing through the opening is the sum of the inflow discharge $P=AV$ and the accelerated discharge due to the moving steep negative wave. The interval Δt must be sufficiently long for the negative wave to develop but not so long as to increase the outflow by the further significant drop in water level of the stretch of reservoir between the outlet and the steep negative wavefront.

The continuity equation between supply and outflow gives

$$Q_0=AV+A_1(C-V)=bk^r, \tag{8}$$

which can be written as [4, p. 213, without last term]

$$Q_0=AV+A_1(C-V)=b\left(k_0+\frac{\alpha(V+U)^2}{2g}\right)^r. \tag{9}$$

For small value of velocity head in comparison with k_0, eq (9) can be expressed in the approximate form

$$Q_0=bk^r=bk_0^r\left[1+\frac{r\alpha(V+U)^2}{2gk_0}\right]. \tag{9a}$$

The second term of [4, p. 213] in form of $(V_0^2/2g)^{3/2}$, which takes here the form

$$-b\left[\frac{\alpha(V+U)^2}{2g}\right]^r \tag{9b}$$

and should be added on the right side of eq (9), can be neglected for reservoir outflows in comparison with the first term. The celerity C is given by eq (2) or (3), and velocity U by eq (5) or (6).

For given h_0, b, and r (values or functions dependent only on the outlet characteristics), and given V and α, the magnitudes A_1, C, U, and $k_0=h_0-h$ are functions of h and of the cross section characteristics near the outflow, at distance $\Delta X=(C-V)\Delta t$. When the cross section at distance ΔX from the outlet is selected, eq (9) enables the computation of the value h. As the discharges around the maximum discharge change slowly (fig. 1), the selection of ΔX is easier. It is, therefore, sufficient to select an average cross section some hundred meters upstream from the outlet. For given h, the values k_0 and U, or A_1 and C can be obtained, so that eq (9) also gives Q_0, the discharge outflow at the time Δt (which is the maximum outflow if ΔX is well selected).

The computation of h is to be made by successive approximations until eq (9) is satisfied for the computed h. Numerical or graphical procedures are feasible for that computation.

A simple graphical procedure for the determination of h, Q_0 and downstream tailwater level H_d is shown in figure 6. Three curves as the functions of h and independent of the height h_0 of the opening are first plotted:

$$Q_1=A_1(C-V)+AV,$$

$$\Delta k_0=\frac{\alpha(V+U)^2}{2g},$$

and

$$Q_3=f(H_d),$$

where Q_3 is the rating curve of the tailwater channel. For the given opening height h_0 the fourth curve $Q_2=bk^r$ is plotted. For some types of outlets, the coefficient b is a function of $k_0/(H_0-h_0)$ and of some other parameters, so that approximate values must be first introduced. For convenience the velocity head Δk_0 is plotted with the same scale as h_0, h, k, and k_0. For an assumed value h or k_0 (point 1) the discharge $Q_1=A_1\ (C-U)+AV$ is obtained (curve 1, point 3), as well as Δk_0 (curve 3, point 2). By adding Δk_0 to k_0 (point 4) the value k is obtained, and $Q_2=bk^r$ (curve 2, point 5). The adjustments have to be made for h until $Q_1=Q_2$. If $Q_1=Q_2$, the value of h is obtained. From that value the discharge

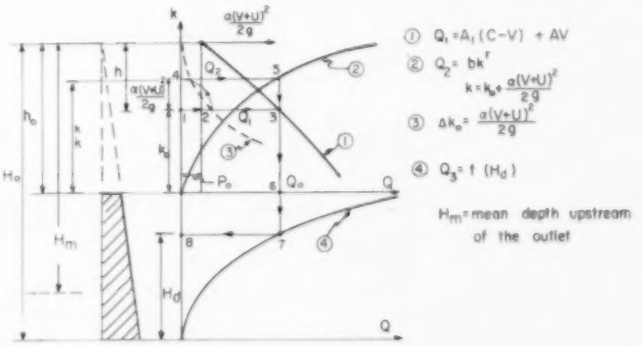


FIGURE 6. *Graphic procedure for the determination of front height of steep negative wave, of maximum discharge, and of maximum tailwater level.*

1—accelerated discharge plus inflow P_0 as function of wavefront height; 2—outflow rating curve; 3—total velocity head as the function of wavefront height; 4—tailwater rating curve.

Q_0 is determined (point 6). For a given Q_0 (curve 4, point 7) the maximum downstream level H_d (point 8) is obtained. When $H_d \leq \beta(H_0 - h_0)$, the outflow is uninfluenced by tailwater level. The coefficient $\beta \geq 1$ is a correction factor which takes into account that, due to the effect of kinetic energy of outflow water, the tailwater levels somewhat higher than $(H_0 - h_0)$ do not influence the outflow.

2.4. Relation of the Height of Steep Negative Waves to the Opening Height of a Rectangular Outlet

For the given opening, the values h and Q_0 depend on the shape of three curves, 1, 2, and 3 of figure 6. The accuracy of curve 2 ($Q = bk^r$) depends on the accuracy of the dimensionless outflow coefficient m which enters into the expression for the coefficient b.

The case of a reservoir, with water at rest, of approximately rectangular cross section $A = B_0 H_m$ near the outlet will be treated here. The bottom of small length of reservoir ΔX can be assumed horizontal, especially under natural conditions with deposited sediments. The simple case of a rectangular sharp-edged opening will be analyzed here as an example, with area Bh_0, where h_0 is much smaller than H_m (fig. 6), and $B \leq B_0$. Equation (9) will be used, and according to [4] and [5], with different symbols used here,

$$m = \frac{Q}{Bh\sqrt{gh}} = \frac{Q}{B_0\sqrt{gh^{3/2}}} \cdot \frac{B_0}{B} = \frac{b}{B_0\sqrt{g}} \cdot \frac{B_0}{B},$$

or

$$m = \frac{B_0}{B} \cdot f\left(\frac{k_0}{H_m - h_0}, R, W\right) \tag{10}$$

with $k_0 = h_0 - h$ and R and W the Reynolds and Weber numbers, respectively.

In this expression B_0 is taken in the denominator instead of the usual B, so that the right side of eq. (10) is multiplied by B_0/B, apart from the contraction effect dependent on B/B_0, in order to facilitate the solution of eq (9), regardless of the changing width B of the opening.

Putting $h_0/H_m = s$ and $h/H_m = p$, the ratio $h/h_0 = p/s$, of the wave height h and the opening height h_0 will be related to s. For the rectangular sharp-edged opening

$$m = \frac{B_0}{B} f\left(\frac{B}{B_0}, \frac{s-p}{1-s}, R, W\right), \tag{11}$$

where m is B_0/B times higher than the usual coefficient m. Other functions of m for other shapes of outlet have to be used, depending on the outlet.

The coefficient m can be thus expressed as a function of dimensionless numbers, so that for each outlet and its characteristics the special relation of eq (11) or of other type can be obtained. For given B, B_0, H_m and h_0, the coefficient m can be expressed as a function of h, or of $p = h/H_m$.

Equation (9) under the above assumptions and in terms of the dimensionless numbers introduced here can be written as

$$s = p + \frac{p^{2/3}}{m^{2/3}}\left(1 - \frac{p}{2}\right)^{1/3} - \frac{\alpha}{4}\frac{p^2(2-p)}{(1-p)^2}. \tag{12}$$

If the function of eq (11) is introduced in eq (12), the relationship $p = f(s)$ is obtained. For the practical solution it is suitable to take m as a parameter and to develop a family of curves for eq (12). Figure 7 gives this family of $p = f(s)$ for $m = 0.25$, 0.50, 0.75, 1.00, 1.25, and 1.50, and figure 8 gives $p/s = f(s)$ for $m = 0.25$, 0.50, 0.75, and 1.00. The families of curves $m = f(p)$ or $m = f\left(\frac{p}{s}\right)$ for given values s are more convenient to use. Both are given in figures 7 and 8 as dashed lines. For these curves $\alpha = 1.05$ was taken as the most applicable constant value.

When eq (11) is determined for a particular case of rectangular opening, with $s = h_0/H_m$ known, the intersection of the curve of eq (11) and the curve $m = f(h/H_m) = f(p)$ or $m = f(h/h_0) = f(p/s)$ from figure 7 or 8 for that s, gives the solution either for $p = h/H_m$ or $p/s = h/h_0$.

If the velocity V in the body of water near the outlet, before the sudden release of additional water, can not be neglected in comparison with the new velocity $(V + U)$, it must be introduced in dimensionless form in eq. (12). In that case, for $w = V_0/\sqrt{gH_m} = V_0/C_m$, with $C_m = \sqrt{gH_m}$, and $V_0 = P/B_0H_m$

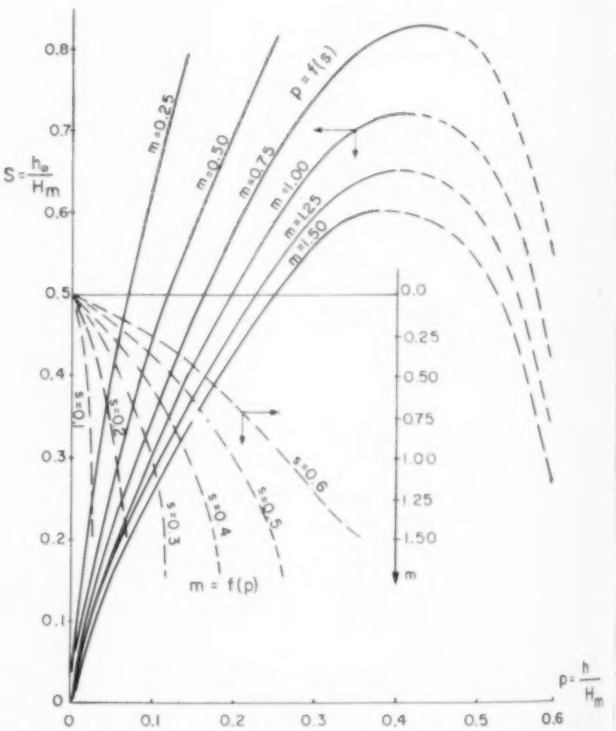


FIGURE 7. *Relation of the relative height of wavefront ($p = h/H_m$) to relative opening height ($s = h_0/H_m$) where H_m is mean depth of reservoir near outlet, with the outflow coefficient m as parameter, for rectangular opening; the second family is the relation of m and p with s as a parameter.*

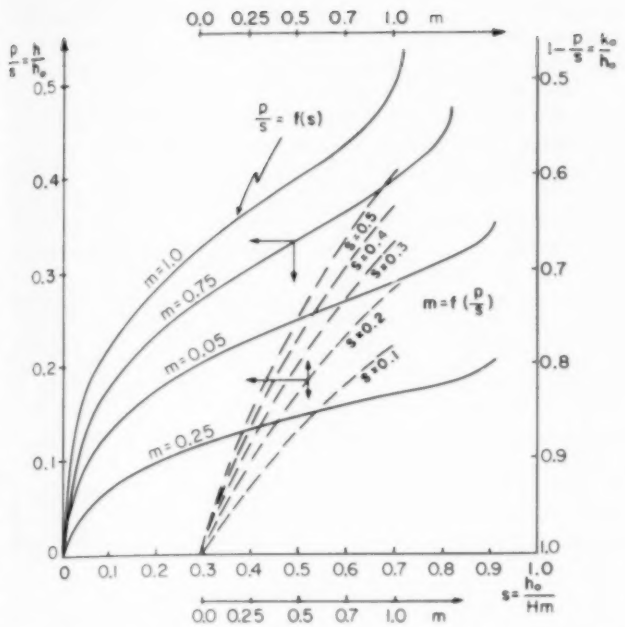


FIGURE 8. *Relation of the ratio of wavefront height and opening height (h/h_0) to relative opening height (h_0/H_m) with outflow coefficient m as parameter for rectangular opening; dashed lines represent the relation of m and p/s with s as parameter.*

$$s=p+\left[\frac{p}{m}\left(1-\frac{p}{2}\right)^{1/2}+\frac{w}{m}\right]^{2/3}-\frac{\alpha}{2}\left[w+\frac{p}{1-p}\sqrt{1-\frac{p}{2}}\right]^{2}. \tag{13}$$

For each given value of w a family of curves $m=f(h/H_m)$ or $m=f(h/h_0)$ can be developed.

Equations (12) and (13) in parametric form can be developed for any regular-shaped outlets (rectangular, triangular, semicircular, circular, trapezoidal, parabolic, semielliptic, etc.). With s, or with s and w, as parameters, the families of curves $m=f(h/h_0)$ or $m=f(h/H_m)$ can be developed once for all cases. With these families available, the determined real function $m=f(h/h_0)$ for a particular outlet of given regular shape will be sufficient for the determination of the height h of the steep negative wave, and of the maximum discharge of outflow hydrograph Q_0 by using eq (8). The nomographs can also be developed for the relations in eqs (12) and (13).

If m is selected to be only the function of h, then eq (12) would give a stable relationship between p and s, or eq (13) would give a family of curves with w as parameter.

2.5. Procedure for the Computation of a Free Outflow Hydrograph

In the case of no inflow into the reservoir, the celerity C depends on the wave height h and on the depth H_0 of reservoir at the outlet before the sudden water release. The celerity is, therefore, the general function

$$C=f\left(x, \frac{h}{H}, \frac{H}{H_0}\right), \tag{14}$$

where $x=1-X/L$ is the relative distance of a cross section from the upstream end of reservoir, X is the distance from the outlet, L is the length of reservoir, H is any possible reservoir depth at the outlet, H_0 is the maximum operational depth, h the height of the steep negative wave.

For a given relative parameter H/H_0, a family of curves $C=f(x, h/H_0)$, can be plotted.

For a large reservoir, H_0 and h will be assumed as the initial or boundary conditions. For a given cross section x the values A, A_1, a, f, C, and U, entering into eqs (2), (3), (5), and (6), are also given.

The general procedure for the second approximation of taking into consideration the effect of the movement of steep negative wave in the case of valley reservoirs is:

(1) To compute the wave height h for given water depth H at the outlet, and to determine the maximum outflow discharge Q_0, as is shown in subsections 2.3 and 2.4, eqs (7) to (13), figures 6 to 8.

(2) To verify that the outflow hydrograph is free, or that the tailwater level is lower than the maximum tailwater level which does not affect the outflow (fig. 6);

(3) To compute the celerity C along the body of water out of given cross section characteristics (A, A_1 and f), eqs (2) and (3);

(4) To compute the velocity V along the body of water for the inflow P during the wave movement, if there is an inflow;

(5) To determine the difference $(C-V)$ along the body of water;

(6) To compute the accelerated discharges by steep negative wave movement along the body of water $P_{w1}=CA_1$ with inflow zero, or $P_{w2}=(C-V)A_1$ with inflow P;

(7) To determine the time, when the wave arrives at the section X and when the acceleration takes place there, by $t=\int_{X=0}^{X}\frac{dx}{C-V}$, or for $1-\frac{X}{L}=x$ (L=length of reservoir) by

$$t=L\int_{x}^{1}\frac{dx}{C-V}, \tag{15}$$

with t_0 the total time of wave travel from the outlet to the upstream end of reservoir (from $x=1$ to $x=0$);

(8) To find the relation between the accelerated discharge and the time of acceleration, $P_w=f(t)$. For given x the pair of values P_w and t gives $P_w=f(t)$.

The integration of P_w over time from $t=0$ to $t=t_0$ gives the volume S_h of the reservoir layer h, determined simply from the storage function $S=f(H)$, figure 10, which can serve to check the function $P_w=f(t)$.

If the inflow during time t_0 cannot be taken as constant (i.e., the natural inflow hydrograph or the release of water from upstream reservoirs are very unsteady), the velocity V at a cross section x is also dependent upon the time as are $C-V$, A_1 and P_w. The inflow can be either a steep or a gradual wave and the determination of $P_w=f(t)$ becomes more complex.

If the inflow is in form of a steep positive or negative wave, the procedure is similar to the above given movement of steep negative wave until two waves traveling in opposite directions meet. The resulting waves will then move in both directions.

In the case where the inflow hydrograph $P=f(t)$ is gradually varied, and where the steep negative wave moves faster upstream than the other negative waves created downstream of the steep wave, a resulting new inflow hydrograph $P_n=f(t)$, composed of $P=f(t)$ and the accelerated discharges by steep negative wave movement, can be determined either by exact or approximate methods. The position of the wave $P=f(t)$ along the reservoir has to be determined. Therefore, the velocity V for given places and the times at which the steep negative wave arrives at those places must be computed in order to obtain the celerity $C-V$. The celerity is now $C=f\left(x,\ t,\frac{h}{H},\frac{H}{H_0}\right)$, with a new variable, t.

The time t_0 of wave travel from one end of the reservoir to the other is the factor which will determine whether or not the inflow hydrograph can be replaced by a constant inflow P_0. Since t_0 for the majority of valley reservoirs is measured in hours, constant inflow can often be assumed during the movement of steep negative wave.

Due to the neglected factors—resistance to the flow, reflection and refraction of the wavefront and transformation of the wavefront due to different celerities of its parts—the computed function $P_w=f(t)$ is an approximation of the true relationship.

The procedure for the second approximation of the free outflow hydrograph would involve the following values and functions:

(1) The outflow head $k=k_0+\frac{\alpha(V+U)^2}{2g}$; (2) the outflow rating curve $Q=b\ k^r$; (3) the characteristics of cross sections along the reservoir, and (4) the total inflow function $P_n=f(t)$, which gives for each time t the discharge $P_n=P+P_w$, and for the corresponding x, also, the place of inflow P_n. The use of the two basic de Saint Venant equations for gradually varied unsteady flow will enable the computation of the free outflow hydrograph in cases where the continuity equation alone can not give reliable results.

In valley reservoirs, where the height of the negative wave is not too large, the use of the continuity equation alone, instead of both continuity and momentum, gives an approximate free outflow hydrograph, sufficiently accurate in the light of background data, assumptions and the needed accuracy of results.

The higher the steep negative wave and the inflow P, the greater is the velocity $(V+U)$, so that the use of the continuity equation alone will give less accurate results. Since the effect of neglected factors in steep negative wave movement is the highest in shallow waters (upstream part of reservoirs) and lowest in the deep waters (lower part of reservoirs) and since P_w is low in the first and high in the second region, the use of only the continuity equation is generally justified. In that case the integration of the storage differential equation (continuity equation) of the type $P-Q=dS/dt$, with P replaced by the new inflow hydrograph $P_n=P+P_w$, can be made by a numerical, graphical, or any other appropriate procedure to obtain the free outflow hydrograph $Q=f(t)$.

Two simple cases will be taken as examples: (*a*) A pyramidal reservoir with rectangular sections (B_X, H_X), with the bottom slope H_0/L, where H_0 is its depth at outlet and L reservoir length, and with B_0 the width of reservoir at the outlet; (*b*) a parabolic section of the type $B=B_0\sqrt{H/H_0}$, where B_0 and H_0 are width and depth respectively of reservoir at the outlet, B and H any width and depth, with the bottom slope H_0/L, and B_x and H_x the surface width and water depth at a cross section X distance from outlet.

2.6. Rectangular Cross Sections

Figure 9 gives the simple shape of the reservoir and the resulting functions along it for the inflow $P=0$. The steep negative wave has the front height h with $p=h/H_0$, and $x=1-(X/L)$, so that $B_x/B_0=H_x/H_0=x$.

The total area of any cross section is $A=x^2B_0H_0$ and the wavefront area $A_1=xhB_0$; $f=h/2$, so that

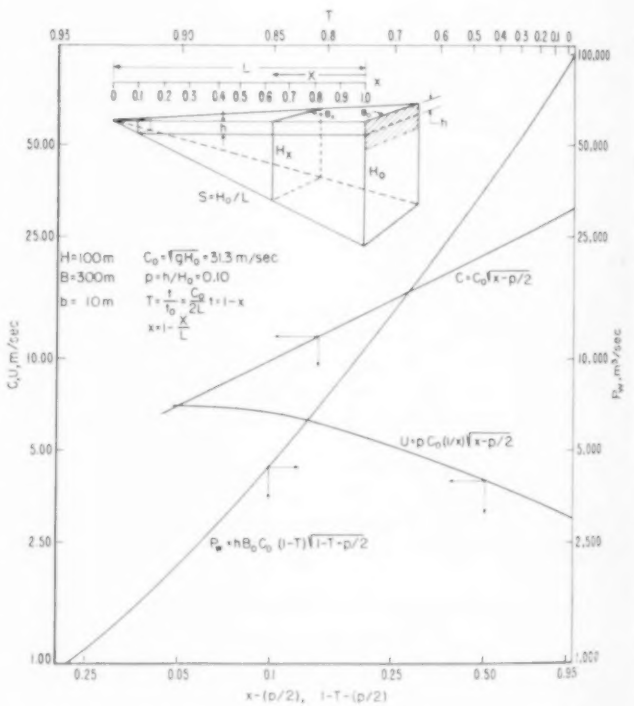


FIGURE 9. *Celerity, velocity of accelerated water, and accelerated discharges along a pyramidal schematic reservoir with rectangular cross sections and rectangular opening of the same width as the reservoir (upper figure).*

from eq. (2)

$$C=C_0\sqrt{x-\frac{p}{2}}, \tag{16}$$

with $C_0=\sqrt{gH_0}$, the maximum celerity of the uppermost part of steep negative wave near the outlet. In the case of constant inflow P_0 and $V=P_0/A$,

$$C-V=C_0\sqrt{x-\frac{p}{2}}-\frac{V_0}{x^2}, \tag{17}$$

with $V_0=P_0/B_0H_0$ the velocity at the outlet before the sudden water release. The new velocity along the body of water after the wave passes is, according to eq (5),

$$V+U=\frac{V_0}{x^2}+C_0\frac{p}{x}\sqrt{x-\frac{p}{2}}. \tag{18}$$

For small values of x eq (16), (17), and (18) have no practical meaning, because C and U are very high, and the resistance plays the major role, so that they should be used for values slightly higher than p. The accelerated discharge plus the inflow P_0 is now $P_n=P_0+P_0$, with

$$P_w=hB_0\left(C_0x\sqrt{x-\frac{p}{2}}-\frac{V_0}{x}\right). \tag{19}$$

The travel time from the outlet to a cross section X distance, in the case $V=0$, is

$$t=\int_0^X\frac{dX}{C}=L\int_x^1\frac{dx}{c}=\frac{2L}{C_0}(1-x). \tag{20}$$

If $x=0$, the travel time along the reservoir for $V_0=0$ is

$$t_0=\frac{2L}{C_0} \tag{21}$$

and $x=1-\frac{t}{t_0}=1-T$, where $T=t/t_0$. With the increase of P_0 the time t_0 increases and the discharge P_w decreases, though $P_n=P_0+P_w$ increases.

The integration of the t—function for $V_0\neq0$ in the form

$$t=L\int_X^1\frac{x^2dx}{C_0x^2\sqrt{x-\frac{p}{2}}-V_0} \tag{22}$$

is best made numerically.

The functions C, U, and P_w are given in logarithmic scales in figure 9 for the following values: $H_0=100$ m; $B_0=300$ m; $h=10$ m; C and U in m/sec and Q in m³/sec.

2.7. Parabolic Cross Section

For given h and H_0, and two cases of the inflow: $P=0$ and $P=P_0$, figure 10 gives the following functions depending on $x=1-(X/L)$ for the parabolic cross sections of the type $B=d\sqrt{H_0H}$, where $d=B_0/H_0$; A; A_1; $a=A/A_1$; C (or $C-V$); U (or $U+V$); $V=P_0/A$; $t=\int_1^x\frac{dx}{C}$ $\left(\text{or } t=\int_1^x\frac{dx}{C-V}\right)$; P_1 (or P_2), accelerated discharge P_w or $P_n=P_0+A_1(C-V)$; and finally P_1 (or P_2) as a function of time t_1 (or t_2).

2.8. Shape of the Hydrograph in the First Moments After Release

A few moments after the release, at the time interval Δt, a full steep negative wave of height h is developed (fig. 1). At that moment the outflow is maximum. The approximate solution for the shape of the rising part of the hydrograph after the sudden water release can be obtained as a straight line for an ideal fluid by using the two-dimensional hydrodynamic equations of motion expressed in Lagrangian form [6].

If the straight line rise is assumed for the total rise, the procedure for the determination of the slope of the hydrograph is simple. The total outflow of water from the time of opening to the time Δt is approximately the volume of water confined between the outflow section and the section at $\Delta X=(C-V)\Delta t$, where $(C-V)$ is the wave celerity at the beginning, and between the starting depth of water H at the outlet and the new depth $H-h$. This volume is ΔS. If it is taken that the maximum discharge Q_0 takes place at time $\Delta t_1<\Delta t$, instead of at Δt, the continuity equation gives $\frac{1}{2}(\Delta t+\Delta t_1)Q_0=\Delta S$, and Δt_1 can be obtained (fig. 1). The mean slope of the straight line for the rising part of the hydrograph is then $Q_0/\Delta t_1$. For small h, the values ΔS and ΔT are very small, and the rise of hydrograph can be taken as vertical.

Figure 11 gives an example of the difference between the outflow hydrograph of the first approximation, curve 1 (emptying of the reservoir with the negative wave movement neglected) and the outflow hydrograph of the second approximation, curve 3 (with the steep negative wave taken into consideration, but with the flow resistance neglected) for the Savage River Reservoir near Cumberland, Md. A rectangular opening (250 m by 18.5 m) with sudden water release in the case of a full reservoir is assumed. Curve 2 gives the fictitious inflow hydrograph computed by the method given above. Though the total volume of the two hydrographs, 1 and 3, is the same, the peak discharge in the second approximation is smaller for about one-third.

3. Effect of Flow Resistance on the Free Outflow Hydrograph

3.1. Two Aspects of the Flow Resistance

The effect of the flow resistance on the free outflow hydrograph depends on the new velocities along the reservoir. The greater the wavefront height h, the less the remaining flow area $A-A_1$ after the steep negative wave has passed, and the greater the inflow and the roughness coefficient in the reservoir, the greater will be the flow resistance effect.

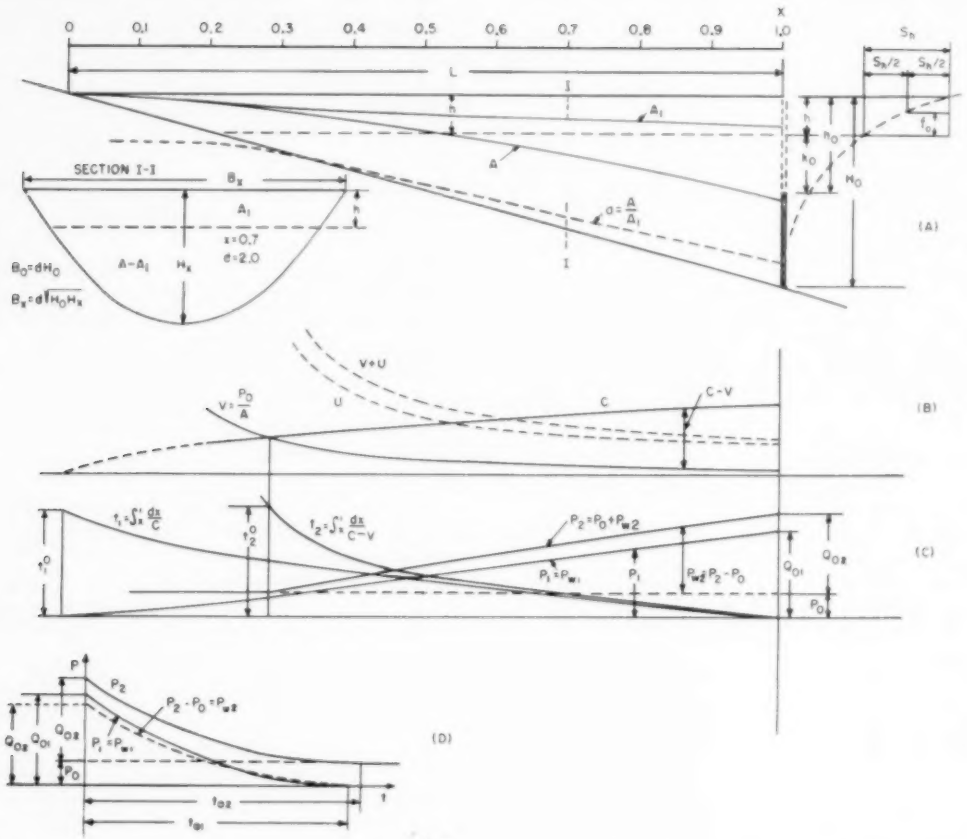


FIGURE 10. *Schematic representation of the cross-section characteristics and hydraulic magnitudes along a pyramidal reservoir with parabolic cross sections.*

(A) reservoir with areas A and A_1 and their ratio $a=A/A_1$ along the reservoir with a cross section at left and storage function of the reservoir at right; (B) celerity C and velocity U of accelerated water, due to the steep wave movement along the reservoir in case where water is at rest, and celerity $C-V$ and velocity $V=P_0/A$ and $V+U$ along the reservoir in case where inflow of water in reservoir is P_0; (C) accelerated discharge and times at which the acceleration take place along the reservoir for two cases: inflow is O and inflow is P_0; (D) fictitious and total inflow hydrographs for reservoir in two cases: inflow is O and inflow is P_0 (third-dashed-curve gives fictitious hydrograph in case inflow $P=O$).

Two aspects of flow resistance can be taken into consideration: (1) Head losses along the reservoir, which can be taken into account by introducing the friction slope into the second or dynamic de Saint Venant equation for unsteady flow gradually varied; and (2) decreasing of front height of the steep negative wave, because of the downstream increase of level due to the resistance head as the wave progresses upstream. Hence, the effective potential energy of wave (energy to be transformed into kinetic energy of accelerated water) will be smaller than the total energy corresponding to the initial height h.

The first aspect of resistance can be taken into account when the routing procedure is based on the two de Saint Venant equations. It is, however, possible to take into consideration the resistance effect on steep negative wave movement when either the continuity equation of the type $P-Q=dS/dt$ or both the continuity and dynamic equations are used for the routing. The effect of flow resistance on the fictitious inflow hydrograph, caused by the steep negative wave movement, will be dealt with in this investigation.

3.2. Effect of the Flow Resistance on the Steep Negative Wave

If a cross section at distance X from the outlet has the following values determined: A, A_1, f, C, V, U, and if the resistance effect is introduced, the height h will decrease along the reservoir, with the slope

$$\frac{dh}{dx}=\frac{n^2(V+U)^2}{R^{4/3}}, \tag{23}$$

where the Manning formula for the flow resistance is used in metric system, where n is the roughness coefficient and R the hydraulic radius of area $(A-A_1)$. At the distance $x=1-(X/L)$ the new wave height h_1 is now

$$h_1=h-h_x=h-\int_1^x \frac{n^2(V+U)^2dx}{R^{4/3}}, \tag{24}$$

where h is the height near the outlet, h_1 the height of

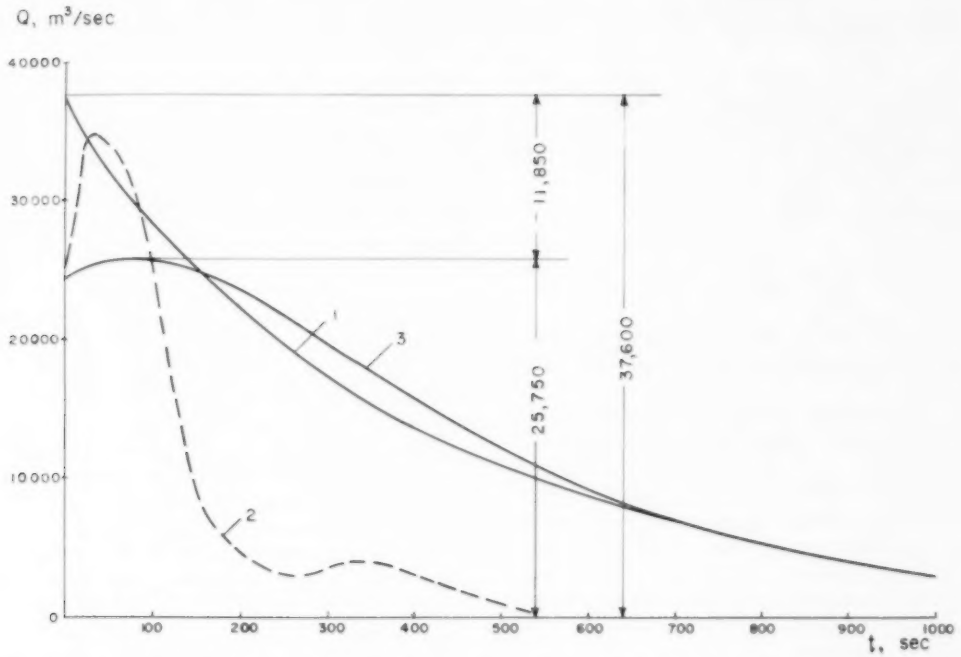


FIGURE 11. *Three hydrographs of the Savage River Reservoir for the rectangular opening 250 m by 18.5 m: 1—first approximation hydrograph, when the movement of all waves is neglected; 2—fictitious inflow hydrograph, when the water accelerated by the movement of steep negative wave is considered as the inflow; 3—second approximation hydrograph, when the movement of steep negative wave is taken into account.*

wave at section x, and h_x the total resistance head from the outlet to the section x. Since U and R are the functions of h_1, the solution has to be made by successive approximations. As h_1 changes slowly, it is often sufficient to start with h—constant for the computation of U and R, and then correct values so obtained by successive approximations.

In fact, the potential energy of the steep negative wave is not all transformed into kinetic energy of accelerated water, one part of it being left as potential energy to be used as the flow resistance head.

The integration of the last term of eq (24) along the reservoir from the outlet to the place where $C-V=0$ would give the total resistance head. The volume S_r which is not accelerated by the wave because of the resistance is part of the total volume S_h (fig. 10) corresponding to wavefront height h, and is

$$S_r=\int_1^{x_0} B_x h_x dx, \qquad (25)$$

where x_0 is the place where $C-V=0$ and B_x is the mean width of the area element for total resistance head h_x at the section x.

The analytical integration of eqs (24) and (25) is not tractable even in simple cases like the pyramidal reservoir of rectangular cross section, figure 9, because the values U and R for a section x depend on the sum of previous dh, and because of the expression $R^{5/3}$.

The practical procedure would be to compute U and R based on constant value h along the reservoir, and to compute the slope of eq (23). By integrating $h_x=\int_1^x \frac{dh}{dx}dx$ the function of total resistance head h_x along the reservoir is obtained. Then $h_1=h-h_x$ gives the effective head of wavefront. With these values of h_x new values U and R are again computed, and eqs (23) and (24) can be obtained as a second approximation. Succeeding approximations can be made by continuing the same procedure.

Due to the fact that the total velocity $(V+U$ increases in the upstream direction while the accelerated discharge decreases, it is usually sufficient to compute h_x to the second or third approximation.

With the new values $h_1=h-h_x=f(x)$ it is now possible to determine a more exact fictitious hydrograph P_w.

For the computed function $h_x=f(x)$ the width B_x has to be determined since $B_x=f(x)$. The integration of eq (25) gives S_r. The product of $B_x h_x$ along the reservoir makes possible the approximate computation of the mean distance f_r of total volume S_r from the level H_0-h:

$$f_r=\frac{E_r}{\gamma S_r}=\frac{1}{2S_r}\int_1^{x_0} B_x h_x^2 dx, \qquad (26)$$

where E_r is the potential energy corresponding to the volume S_r and the resistance head and γ is the unit weight of water.

The energy of the steep negative wave (fig. 10) is

$$E_w=\gamma f_0 S_h, \tag{27}$$

where S_h is total volume of reservoir between depths H_0-h and H_0, and f_0 is the distance from the center of gravity of S_h to the H_0-h (fig. 10).

The ratio of energies

$$e=\frac{E_r}{E_w}=\frac{f_r S_r}{f_0 S_h} \tag{28}$$

can be taken as a measure of the flow resistance effect on the transformation of the front height of steep negative wave.

As h_x increases somewhat faster than B_x decreases in the upstream direction, $B_x h_x$ increases from the outlet upstream. The effect of flow resistance decreases the accelerated flow P_w, slower at its highest values near the outlet and faster at its lowest values at the upstream end of the reservoir. The fictitious inflow hydrograph will be now closer to a straight line. As a practical approximation for rapid computations this hydrograph can be taken as triangular-shaped for a valley reservoir.

3.3. Simplest Procedure for the Computation of the Steep Negative Wave Effect With Resistance Influence

A simple procedure for the computation of the approximate free outflow hydrograph in this case would be:

(1) Compute h and Q_0; (2) compute A, A_1, f, C, V, U along the reservoir for constant h; (3) determine h_x along the reservoir; (4) recompute all values under (2); (5) recompute h_x and all values under (2) until a sufficient approximation of the true values is obtained; (6) compute B_x as $f(x)$; (7) compute S_r for given h_x and B_x curves; (8) determine S_h from the storage function; (9) determine the difference S_h-S_r; (10) determine the triangular shape of fictitious inflow hydrograph P_w with the maximum discharge Q_0 and the total time of fictitious hydrograph.

$$t_0=\frac{2(S_h-S_r)}{Q_0}. \tag{29}$$

The sum of the original and of the fictitious inflow hydrograph gives the total inflow hydrograph. The new inflow hydrograph, the maximum discharge Q_0 at the new starting level k (fig. 6), the outflow rating curve and the storage function are now sufficient elements for any suitable routing procedure, based on the continuity equation only, of obtaining the free outflow hydrograph.

If the routing procedure involves the use of both the continuity and dynamic equations for unsteady flow which is gradually varied, the computed fictitious inflow discharges P_w, with their positions in time and along the reservoir, and the corresponding reservoir depths H_0-h_1 for the same times and places are the boundary conditions to be introduced in those equations.

4. Discussion

The computed outflow hydrographs from valley reservoirs in the case of sudden water release can be only approximations of the true hydrographs. The shape and the resistance to flow of valley reservoirs as well as the water movement itself along them are usually very complex. The closer an approximated hydrograph is to the true hydrograph, the more work is involved in the computation. The digital computers will, in some cases, give the most accurate results. The use of computers, however, is justified only in cases where very reliable and sufficient records and data about the reservoir, outlet, and inflow are available and where the great precision of results and speed of computation make their use economically feasible. Therefore, the use of simple numerical or graphical procedures for the computation of approximate outflow hydrographs is often very convenient, sensible, and economically desirable.

The movement of the steep negative wave along the reservoir accelerates water in the reservoir. Many factors affect this acceleration so that the approximations of the true outflow hydrograph depend on the factors either taken into account or neglected. The accelerated discharges are considered here as a fictitious inflow hydrograph to be taken as given either when only the storage continuity equation $P-Q=dS/dT$ is used for routing procedure or when both the continuity and dynamic equations are applied. The degree of unsteadiness of water movement is the criterion for the use of only the continuity equation or of both it and the momentum equation for the routing. The greater the height of the steep negative wave in relation to the reservoir depths, the higher and more unsteady the inflows, and the greater the flow resistance, then the higher will be the degree of unsteadiness of flow motion, and the use of the two de Saint Venant equations with the computation involved will be more necessary. The development of expressions for the quantitative determination of this criterion is a special problem not treated here. A criterion for the effect of flow resistance on the movement of steep negative wave is given by eq (28). When the ratio e is higher than a selected limit, the flow resistance effect has to be taken into account, regardless of the routing procedure.

The fact that the steep wave is a discontinuity, not suitable to be considered as an unsteady flow gradually varied, and the fact that the movement and transformation of steep negative wave is not affected by small downstream disturbances in the reservoir, make the treatment of sudden water release different from an ordinary routing procedure. The determination of the outflow hydrograph can be divided into two procedures: (*a*) The effect of steep negative wave and of flow resistance on the water acceleration, combined with the inflow hydrograph during the period of wave movement; (*b*) the routing procedures, based on the position in place and time of accelerated discharges and inflow discharges along the reservoir, as boundary conditions.

As the maximum outflow discharge Q_0 in the case of sudden water release of relatively larger openings is significantly lower than the discharge Q_{max} which corresponds to the given reservoir level and opening with already accelerated water in the reservoir, the maximum tailwater level for Q_0 is much lower than for Q_{max}, so that the free outflow hydrograph (outflows without tailwater effect) is much more common than one would expect.

WASHINGTON, D.C. (Paper 63B2–13)

6. References

[1] M. J. Boussinesq, Essai sur la théorie des eaux courantes, Mémoires presentés à l'Academie des sciences de l'Institut de France **[2] 23** (Paris, 1877).

[2] H. Favre, Etude théoretique et expérimentale des ondes de translation dans les canaux découverts (Dunod, Paris, 1935).

[3] V. M. Yevdjevich, Analytical integration of the differential equation for water storage, J. Research NBS **63B**, 43 (1959).

[4] I. W. Howe, Flow measurement (ch. 3, Engineering hydraulics, Hunter Rouse, ed., pp. 212–222, John Wiley & Sons, Inc., New York, N.Y., 1950).

[5] C. E. Kindvater and R. W. Carter, Discharge characteristics of rectangular thin-plate weirs, J. Hydraulics Div., Proc. Am. Soc. Civil Engrs. (Dec. 1957) Paper 1453.

[6] F. V. Pohle, Motion of water due to breaking of a dam, and related problems, symposium on gravity waves, NBS Circ. 521 (1952).

Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 63B, No. 2, October-December 1959

Uniform Asymptotic Expansions for Weber Parabolic Cylinder Functions of Large Orders

F. W. J. Olver*

(April 6, 1959)

Expansions of solutions of the differential equation

$$\frac{d^2w}{dt^2}=\mu^4(t^2-1)w,$$

are sought for large values of $|\mu|$, which are uniformly valid with respect to arg μ and unrestricted values of the complex variable t. Two types of expansion are found. Those of the first type are in terms of elementary functions and are valid outside the neighborhoods of the points $t=\pm 1$, the turning points of the differential equation. The second are in terms of Airy functions and hold in unbounded regions containing one of the turning points.

The special forms of the expansions when the variables are real are considered in detail, and asymptotic expansions for the zeros of solutions of the differential equation are found by reversion. Numerical examples are included.

Contents

	Page
Part 1. Introduction	131
1. Introduction and general summary	131
2. Relevant properties of parabolic cylinder functions and the gamma function	132
Part 2. Expansions in terms of elementary functions	135
3. Asymptotic solutions of the differential equation	135
4. Identification of solutions	140
5. Use of connection formulas	141
6. Summary of expansions in terms of elementary functions	146
Part 3. Expansions in terms of Airy functions	148
7. Asymptotic solutions of the differential equation	148
8. Identification of solutions	151
9. Use of connection formulas	153
10. Summary of expansions in terms of Airy functions	155
Part 4. Results for real variables	156
11. The equation $w''=(\frac{1}{4}x^2+a)w$	156
12. The equation $w''=(a-\frac{1}{4}x^2)w$	159
13. Numerical examples	162
Part 5. Zeros and associated values	164
14. Zeros of $U(a,z)$, $U'(a,z)$, $\overline{U}(a,z)$, and $\overline{U'}(a,z)$	164
15. Zeros of $W(a,x)$ and $W'(a,x)$	166
16. Numerical examples	168
References	169

Part 1. Introduction

1. Introduction and General Summary

The purpose of this paper is to determine asymptotic expansions of solutions of the differential equation

$$\frac{d^2w}{dt^2}=\mu^4(t^2-1)w \tag{1.1}$$

for large values of the complex parameter μ which are uniformly valid with respect to arg μ and the complex variable t.

The immediate reason for this investigation is that the solutions of eq (1.1) are basic functions in the theory of the asymptotic solution of the differential equation

$$\frac{d^2w}{dt^2}=\{\mu^4p(t)+q(t)\}w \tag{1.2}$$

for large values of $|\mu|$ in a region in which $p(t)$ has two simple zeros, just as Airy functions are

*Present address: National Physical Laboratory, Teddington, Middlesex, England.

basic functions in the theory of eq (1.2) in a region containing one simple zero of $p(t)$. A knowledge of the uniform asymptotic behavior of the solutions of (1.1) is required in developing the asymptotic theory of (1.2).

Equation (1.1) is of interest, however, in other connections, for example in the study of wave motion inside or outside parabolic cylinders, and wave propagation in an inhomogeneous atmosphere. Hermite polynomials of order n are expressible in terms of the solutions of the equation when $\mu^2=2n+1$. Accordingly, for the sake of completeness many results are derived in this paper which go beyond those needed for the immediate purpose stated in the preceding paragraph.

The desired asymptotic expansions are obtained by application of the general theory of the asymptotic solution of linear differential equations of the second order as developed by the present writer in [9].[1] From the standpoint of the general theory, eq (1.1) is characterized by having turning points at $t=\pm 1$. Asymptotic expansions in terms of elementary functions can be found in any region bounded away from these points and expansions of this kind are obtained in part 2 by application of theorem A of [9].

In order to determine the behavior near the turning points, expansions in terms of Airy functions need to be used. In part 3 these expansions are obtained by application of theorem B of [9]; they are valid in unbounded t-regions which include one (but not both) of the turning points. All combinations of μ and t are covered.

When the variables are real, eq (1.1) has important special solutions. Expansions of these solutions are derived in part 4 from the results of parts 2 and 3. Also included in part 4 are numerical examples illustrating the powerful nature of the expansions.

In part 5 uniform asymptotic expansions for the zeros and associated values are obtained by reversion and illustrated by numerical examples. Much of the analysis of this part and some in earlier parts is similar to that used in [8] in the theory of Bessel functions of large order, and advantage is taken of this analogy whenever possible.

Investigations of the asymptotic solution of eq (1.1) for large $|\mu|$ have been made by several writers. Their work on expansions in terms of elementary functions is described in section 6 of part 2, and on expansions in terms of Airy functions in section 10 of part 3. At this stage it suffices to remark that the expansions obtained in this paper include practically all of the earlier results and extend them in various ways.

2. Relevant Properties of Parabolic Cylinder Functions and the Gamma Function

A comprehensive account of the properties of parabolic cylinder functions is given by Miller in the Introduction to [4]. For the purpose of reference we collect here the properties which will be required in this paper. We use Miller's notation rather than that of Whittaker [13] and adopt Miller's choice of solutions when the variables are real.

The standard form of differential equation for the parabolic cylinder functions is

$$\frac{d^2w}{dz^2}=\left(\frac{1}{4}z^2+a\right)w. \tag{2.1}$$

The principal solution $U(a,z)$ is determined by the condition

$$U(a,z)\sim z^{-a-\frac{1}{2}}e^{-\frac{1}{4}z^2} \qquad \text{as } z\to+\infty. \tag{2.2}$$

It is an integral (entire) function of z and an integral function of a. In terms of Whittaker's notation $D_n(z)$ for parabolic cylinder functions and $W_{k,m}(z)$ for the confluent hypergeometric function, we have

$$U(a,z)=D_{-a-\frac{1}{2}}(z)=2^{-\frac{1}{2}a}z^{-\frac{1}{2}}W_{-\frac{1}{2}a,-\frac{1}{4}}\left(\frac{1}{2}z^2\right). \tag{2.3}$$

[1] Figures in brackets indicate the literature references at the end of this paper.

Other solutions of (2.1) are $U(a,-z)$, $U(-a,iz)$ and $U(-a,-iz)$. The connection formulas are

$$U(-a,\pm iz)=(2\pi)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}+a\right)\{e^{-i\pi(\frac{1}{2}a-\frac{1}{4})}U(a,\pm z)+e^{i\pi(\frac{1}{2}a-\frac{1}{4})}U(a,\mp z)\}, \tag{2.4}$$

$$U(a,\pm z)=(2\pi)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}-a\right)\{e^{-i\pi(\frac{1}{2}a+\frac{1}{4})}U(-a,\pm iz)+e^{i\pi(\frac{1}{2}a+\frac{1}{4})}U(-a,\mp iz)\}. \tag{2.5}$$

For fixed a and large $|z|$

$$U(a,z)=z^{-a-\frac{1}{2}}e^{-\frac{1}{4}z^2}\{1+O(|z|^{-2})\} \qquad \left(|\arg z|\leq\frac{3}{4}\pi-\epsilon\right), \tag{2.6}$$

where[2] the O is uniform with respect to arg z. Here and elsewhere ϵ denotes an arbitrary positive number which is independent of all other variables. From (2.4) and (2.6), we obtain

$$U(a,-z)=e^{-i\pi(a+\frac{1}{2})}z^{-a-\frac{1}{2}}e^{-\frac{1}{4}z^2}\{1+O(|z|^{-2})\}+\frac{(2\pi)^{\frac{1}{2}}}{\Gamma(\frac{1}{2}+a)}\,z^{a-\frac{1}{2}}e^{\frac{1}{4}z^2}\{1+O(|z|^{-2})\}, \tag{2.7}$$

valid when $-\frac{3}{4}\pi+\epsilon\leq\arg z\leq\frac{1}{4}\pi-\epsilon$. The same result with the sign of i changed holds when $-\frac{1}{4}\pi+\epsilon\leq\arg z\leq\frac{3}{4}\pi-\epsilon$; the apparent discrepancy between the formulas in their common region of validity is merely an example of the Stokes phenomenon.

Let the z-plane be divided into four sectors $\mathbf{M}_0$, $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$, defined by $|\arg z|\leq\frac{1}{4}\pi$, $\frac{1}{4}\pi\leq\arg z\leq\frac{3}{4}\pi$, $|\arg(-z)|\leq\frac{1}{4}\pi$ and $-\frac{3}{4}\pi\leq\arg z\leq-\frac{1}{4}\pi$, respectively. Then eq (2.6) shows that for large $|z|$, $U(a,z)$ is exponentially small in $\mathbf{M}_0$ and exponentially large in $\mathbf{M}_1$ and $\mathbf{M}_3$. Equation (2.7) shows that $U(a,z)$ is also exponentially large in $\mathbf{M}_2$, unless a has one of the values $-\frac{1}{2}$, $-\frac{3}{2}$, $-\frac{5}{2}$, . . . , in which case it is exponentially small. In fact in this event

$$U(-n-\tfrac{1}{2},-z)=(-)^nU(-n-\tfrac{1}{2},z) \qquad (n=0,1,2,\ \ldots\ ). \tag{2.8}$$

The solutions $U(a,\pm z)$, $U(-a,\pm iz)$ clearly form a numerically satisfactory set in the complex plane, because in each sector $\mathbf{M}_j$ one of the solutions is exponentially small and at least two others exponentially large. Accordingly, we may expect that a knowledge of the asymptotic behavior of $U(a,z)$ for large $|a|$ in the sector $|\arg z|\leq\frac{3}{4}\pi$, or even the half-plane $|\arg z|\leq\frac{1}{2}\pi$, will enable us to determine the behavior of any solution of (2.1) over the whole z-plane by use of connection formulas.

Solutions when the variables are real. There are two forms of eq (2.1) of importance when the variables are real, namely

$$\frac{d^2w}{dx^2}=\left(\frac{1}{4}x^2+a\right)w, \tag{2.9}$$

and

$$\frac{d^2w}{dx^2}=\left(a-\frac{1}{4}x^2\right)w, \tag{2.10}$$

in which a and x are real. Equation (2.10) is obtained from (2.1) by setting $z=xe^{-\frac{1}{4}\pi i}$ and replacing a by ia.

Real solutions of (2.9) are $U(a,x)$ and $U(a,-x)$ but they do not form a satisfactory pair for all values of a; in particular they are linearly dependent when $a=-\frac{1}{2}$, $-\frac{3}{2}$, $-\frac{5}{2}$, ... The fundamental solutions are taken to be $U(a,x)$ and $V(a,x)$, where

$$V(a,x)=(1/\pi)\Gamma\left(\frac{1}{2}+a\right)\{\sin\pi a\cdot U(a,x)+U(a,-x)\}. \tag{2.11}$$

---

[2] Miller [4, p. 27], who is more concerned with asymptotic expansions fulfilling the uniqueness conditions of Watson [11] than the Poincaré condition, gives the range $|\arg z|<\frac{1}{2}\pi$. The validity, in Poincaré's sense, for the wider range is established in [14, p. 347].

For all values of a, $U(a,x)$ and $V(a,x)$ are linearly independent. When a is negative it is often convenient to use in place of $V(a,x)$ its multiple $\overline{U}(a,x)$, given by

$$\overline{U}(a,x)=\Gamma\left(\frac{1}{2}-a\right)V(a,x)=\tan\,\pi a\cdot U(a,x)+\sec\,\pi a\cdot U(a,-x). \tag{2.12}$$

When $a+\frac{1}{2}$ is a positive integer $\overline{U}(a,x)$ is infinite.

Wronskian relations for these solutions are

$$/\!/\,\{U(a,x),V(a,x)\}=(2/\pi)^{\frac{1}{2}},\qquad /\!/\,\{U(a,x),\overline{U}(a,x)\}=(2/\pi)^{\frac{1}{2}}\Gamma\left(\frac{1}{2}-a\right). \tag{2.13}$$

The Hermite polynomials

$$H_n(x)=(-)^n e^{x^2}\frac{d^n}{dx^n}e^{-x^2} \tag{2.14}$$

are related to the function $U(a,x)$ by

$$H_n(x)=2^{\frac{1}{2}n}e^{\frac{1}{2}x^2}U\left(-n-\frac{1}{2},x\sqrt{2}\right). \tag{2.15}$$

For eq (2.10), the fundamental pair of real solutions is taken to be $W(a,x)$ and $W(a,-x)$, defined by

$$W(a,x)=(2k)^{\frac{1}{2}}e^{\frac{1}{4}\pi a}\,\mathrm{Re}\{e^{i(\frac{1}{2}\phi_2+\frac{1}{8}\pi)}U(ia,xe^{-\frac{1}{4}\pi i})\}, \tag{2.16}$$

$$W(a,-x)=(2/k)^{\frac{1}{2}}e^{\frac{1}{4}\pi a}\,\mathrm{Im}\{e^{i(\frac{1}{2}\phi_2+\frac{1}{8}\pi)}U(ia,xe^{-\frac{1}{4}\pi i})\}. \tag{2.17}$$

where

$$k=(1+e^{2\pi a})^{\frac{1}{2}}-e^{\pi a},\qquad 1/k=(1+e^{2\pi a})^{\frac{1}{2}}+e^{\pi a}\qquad \phi_2=\arg\,\Gamma\left(\frac{1}{2}+ia\right). \tag{2.18}$$

The value of $\arg\,\Gamma(\frac{1}{2}+ia)$ here is not the principal one but is determined by the conditions that it is a continuous function of a and equals $a\,\ln a-a+O(a^{-1})$ as $a\to+\infty$.

The Wronskian relation is

$$/\!/\,\{W(a,x),W(a,-x)\}=1. \tag{2.19}$$

Gamma function expansions. We shall need the asymptotic expansion of the function $\Gamma\left(\frac{1}{2}+z\right)$ and its reciprocal for large $|z|$. From Stirling's series [5, chapter 9], we have

$$\ln\,\Gamma\left(\frac{1}{2}+z\right)=\frac{1}{2}\ln 2\pi+z\ln z-z+Z(z), \tag{2.20}$$

where

$$Z(z)\sim\sum_{s=1}^{\infty}\frac{B_{2s}(\frac{1}{2})}{2s(2s-1)}\frac{1}{z^{2s-1}}\equiv-\frac{1}{24}\frac{1}{z}+\frac{7}{2880}\frac{1}{z^3}-\frac{31}{40320}\frac{1}{z^5}+\ldots\qquad(|\arg z|\leq\pi-\epsilon) \tag{2.21}$$

uniformly with respect to $\arg z$, B_{2s} denoting the Bernoulli polynomial of degree $2s$. Hence

$$\Gamma\left(\frac{1}{2}+z\right)\sim(2\pi)^{\frac{1}{2}}e^{-z}z^z\sum_{s=0}^{\infty}\frac{\gamma_s}{z^s}\qquad(|\arg z|\leq\pi-\epsilon), \tag{2.22}$$

where the constants are determined by the asymptotic identity

$$\sum_{s=0}^{\infty}\frac{\gamma_s}{z^s}\sim\exp\{Z(z)\}, \tag{2.23}$$

or equivalently,

$$\sum_{s=1}^{\infty}\frac{s\gamma_s}{z^{s+1}}\sim\left\{\sum_{s=1}^{\infty}\frac{B_{2s}(\frac{1}{2})}{2s}\frac{1}{z^{2s}}\right\}\left\{\sum_{s=0}^{\infty}\frac{\gamma_s}{z^s}\right\}. \tag{2.24}$$

Numerical calculation yields

$$\gamma_0=1, \qquad \gamma_1=-\frac{1}{24}, \qquad \gamma_2=\frac{1}{1152}, \qquad \gamma_3=\frac{1003}{4\,14720}, \qquad \gamma_4=-\frac{4027}{398\,13120}. \tag{2.25}$$

Since $Z(z)$ is an odd function of z, it follows from (2.23) that

$$\left\{\sum_{s=0}^{\infty}\frac{\gamma_s}{z^s}\right\}\left\{\sum_{s=0}^{\infty}(-)^s\frac{\gamma_s}{z^s}\right\}\sim 1, \tag{2.26}$$

identically, and hence that

$$\frac{1}{\Gamma(\frac{1}{2}+z)}\sim\frac{e^z z^{-z}}{(2\pi)^{\frac{1}{2}}}\sum_{s=0}^{\infty}(-)^s\frac{\gamma_s}{z^s} \qquad (|\arg z|\leq\pi-\epsilon). \tag{2.27}$$

Part 2. Expansions in Terms of Elementary Functions

3. Asymptotic Solutions of the Differential Equation

In order to bring the differential eq (1.1) into the standard form required for the application of theorem A of [9], we take new variables $\xi=\xi(t)$ and W (not the $W(a,x)$ of sec. 2), defined by

$$\left(\frac{d\xi}{dt}\right)^2=t^2-1, \qquad W=(t^2-1)^{\frac{1}{4}}w \tag{3.1}$$

(cf. [7] sec. 2). Equation (1.1) then becomes

$$\frac{d^2W}{d\xi^2}=\{\mu^4+\mathscr{F}(\xi)\}W, \tag{3.2}$$

where

$$\mathscr{F}(\xi)=(t^2-1)^{-\frac{1}{4}}\frac{d^2}{d\xi^2}(t^2-1)^{\frac{1}{4}}=-\frac{3t^2+2}{4(t^2-1)^3}. \tag{3.3}$$

For convenience we prescribe that $\xi=0$ when $t=1$. From the first of eqs (3.1) we then obtain

$$\xi=\int_1^t(t^2-1)^{\frac{1}{2}}\,dt=\tfrac{1}{2}t(t^2-1)^{\frac{1}{2}}-\tfrac{1}{2}\ln\{t+(t^2-1)^{\frac{1}{2}}\}, \tag{3.4}$$

and on expansion

$$\xi=\tfrac{1}{2}t^2-\tfrac{1}{2}\ln 2t-\tfrac{1}{4}+\tfrac{1}{4}\sum_{s=1}^{\infty}\frac{1.3.5\ldots(2s-1)}{s(s+1)!(2t^2)^s} \qquad (|t|>1). \tag{3.5}$$

Temporarily we suppose that the branches of the many-valued functions occurring in (3.4) and (3.5) are positive when $t>1$ and determined by continuity elsewhere. The relation

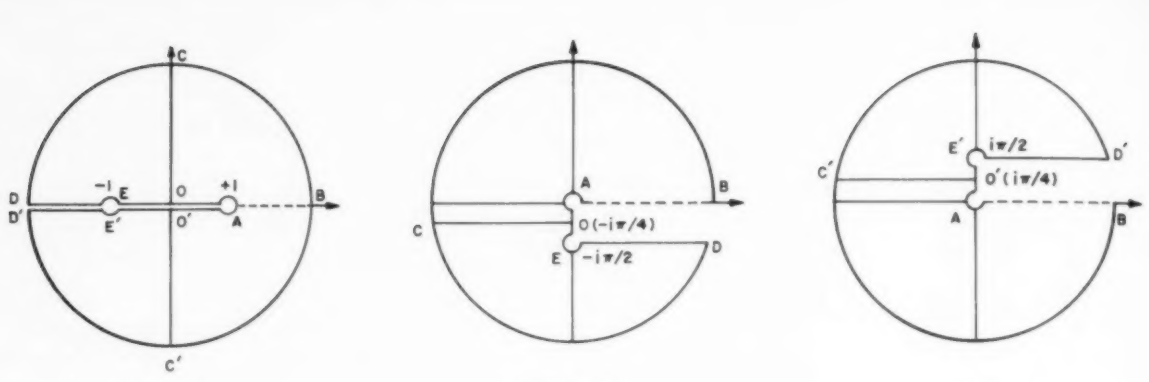
$$\xi(te^{\pm i\pi})=\xi(t)\mp\tfrac{1}{2}i\pi \tag{3.6}$$

follows immediately from (3.5) when $|t|>1$, and thence by analytic continuation throughout the t-plane cut along the join of $t=-1$ and $t=+1$.

The mapping of the t-plane, cut along the real axis from $-\infty$ to $+1$, is illustrated in figure 1. The center diagram corresponds to the upper half of the t-plane and the right-hand diagram to the lower half. They are of course conjugate to each other. Features of the mapping are that the upper side of the real t-axis is mapped on three straight lines BA (Im $\xi=0$, Re $\xi\geq 0$), AE (Re $\xi=0$, $0\geq$ Im $\xi\geq-\frac{1}{2}\pi$), ED (Im $\xi=-\frac{1}{2}\pi$, Re $\xi\geq 0$), and that the positive imaginary t-axis is mapped on the straight line OC (Im $\xi=-\frac{1}{4}\pi$, Re $\xi\leq 0$).

Next, we make the substitutions (cf. [9], sec. 1)

$$u=|\mu|^2, \qquad \theta=\arg\mu, \qquad \xi=e^{-2i\theta}x \tag{3.7}$$



t-plane *ξ-plane (i)* *ξ-plane (ii)*

FIGURE 1. *Mapping of the function* $\xi(t)$.

(the x here is not to be confused with the real variable of sec. 2). Then (3.2) becomes

$$\frac{d^2W}{dx^2}=\{u^2+e^{-4i\theta}\mathscr{F}(e^{-2i\theta}x)\}W, \tag{3.8}$$

which has the form of eq (2.1) of [9]. The relationship between x and t is given by

$$x=x(t)=e^{2i\theta}\xi(t)=e^{2i\theta}\int_1^t (t^2-1)^{\frac{1}{2}}dt; \tag{3.9}$$

the x-map of the t-plane is obtained from the ξ-map by rotation through an angle 2θ. The function $\mathscr{F}$ is expressed in terms of t by (3.3). For large $|x|$ we deduce from (3.3) and (3.5) that

$$e^{-4i\theta}\mathscr{F}(e^{-2i\theta}x)\sim-\frac{3}{4}e^{-4i\theta}t^{-4}\sim-\frac{3}{16x^2}, \tag{3.10}$$

uniformly with respect to bounded arg x and bounded θ.

In consequence of the last result we could, if we wished, take the x-domain $\mathbf{D}(\theta)$ of [9], section 2 to be the map of the t-plane cut along the real axis from $-\infty$ to $+1$ with the circles $|t\pm1|<\delta'$ removed, δ' being a fixed small positive number. The preliminary condition (2.2) of [9] would then be satisfied with $\sigma=1$. Greater regions of validity in the desired expansions will result however, if we arrange that the boundaries of $\mathbf{D}(\theta)$ are parallel to the positive imaginary x-axis where possible. Accordingly, we shall now investigate what we shall call the *principal curves* in the t-plane, namely the t-maps of the straight lines in the ξ-plane passing through A ($\xi=0$) and E ($\xi=\mp\frac{1}{2}i\pi$) and inclined at an angle -2θ to the imaginary ξ-axis. In other words, the principal curves are the level curves of the function $e^{x(t)}$ which pass through $t=\pm1$.

Near $t=1$ we find from (3.9) that

$$x(t)\sim\frac{1}{3}2^{\frac{1}{2}}e^{2i\theta}(t-1)^{\frac{3}{2}}. \tag{3.11}$$

Hence three principal curves emanate from $t=1$, and have as tangents the rays

$$\arg(t-1)=\pm\frac{1}{3}\pi-\frac{4}{3}\theta,\quad \pi-\frac{4}{3}\theta.$$

Equation (3.6) shows that three principal curves also emanate from $t=-1$.

For large $|t|$ we have from (3.5)

$$x(t)\sim\frac{1}{2}e^{2i\theta}t^2. \tag{3.12}$$

Hence the six principal curves are asymptotic to the four rays $\arg t=\pm\frac{1}{4}\pi-\theta$, $\pm\frac{3}{4}\pi-\theta$.

Figure 2 illustrates the principal curves when $0\leq\theta<\pi$. Those through A are the continuous curves AP, AQ, AS when $0<\theta<\frac{1}{2}\pi$, and AP, AQ, AR when $\frac{1}{2}\pi<\theta<\pi$. When $\theta=0$, AQ degenerates into the curve EQ plus the join AE, and when $\theta=\frac{1}{2}\pi$, AS, AR degenerate into AE+ES, AE+ER, respectively. The principal curves through E are the images in the origin of those through A; this can be verified by using the continuation formula (3.6).

We can now define the branch of the many-valued function $\xi(t)$ which will be the most convenient for our purposes. When $0\leq\theta<\pi$ we cut the t-plane along the principal curves AQ and ER and select the branch of $\xi(t)$ which is continuous in the cut plane and for which

$$\xi(0)=\frac{1}{4}i\pi \quad (0<\theta<\pi), \qquad \xi(0-0i)=\frac{1}{4}i\pi \quad (\theta=0). \tag{3.13}$$

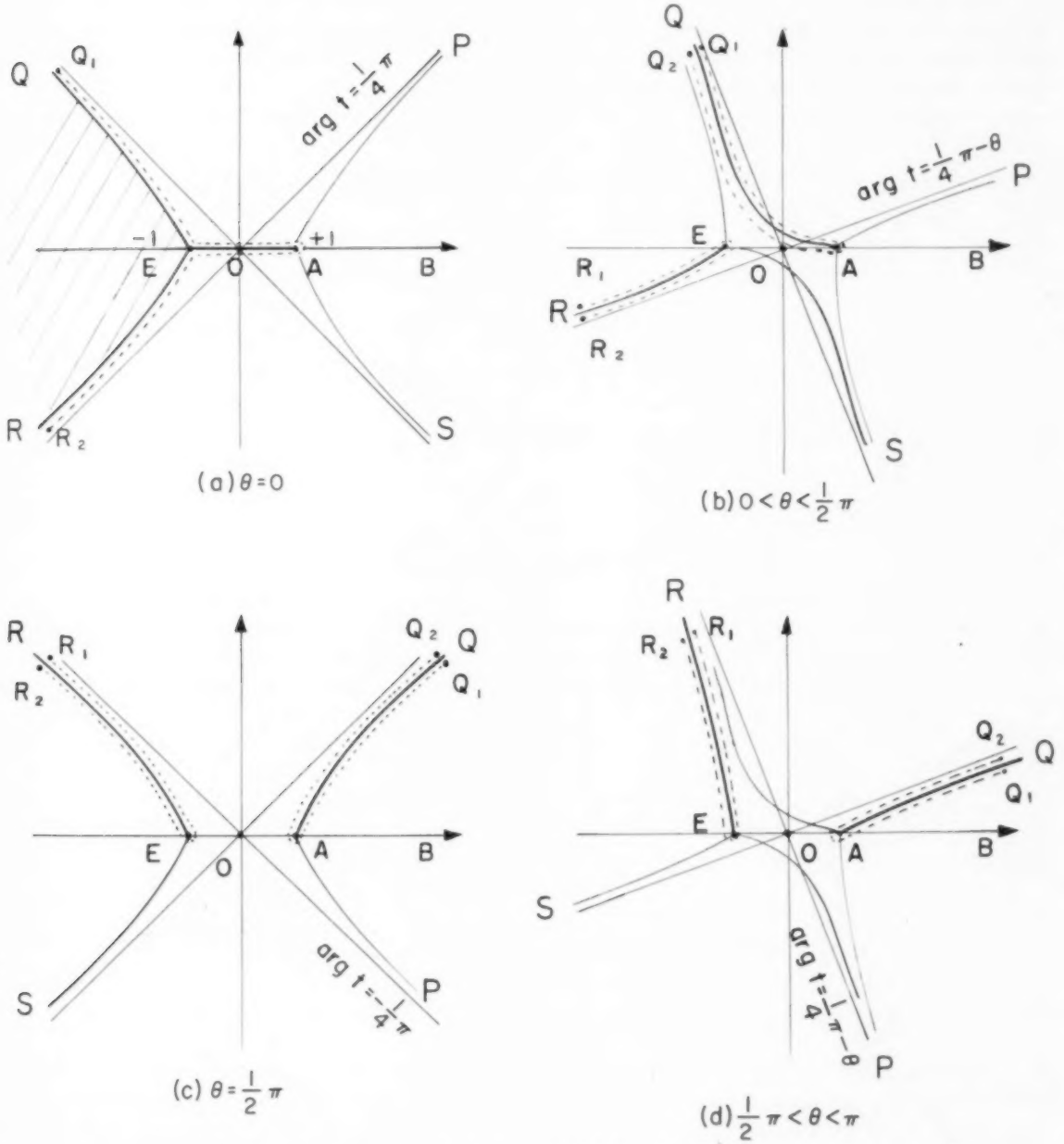


FIGURE 2. *t-plane: principal curves and domains* $\mathbf{S}(\theta)$.

When $-\pi<\theta\leq 0$ the cuts are taken to be the conjugates of those for $-\theta$, and the branch is fixed by

$$\xi(0)=-\frac{1}{4}i\pi \quad (-\pi<\theta<0), \qquad \xi(0+0i)=-\frac{1}{4}i\pi \quad (\theta=0). \tag{3.14}$$

It is evident from (3.6) that (3.13) and (3.14) are consistent when $\theta=0$. *Henceforth we understand ξ and $\xi(t)$ to denote this branch.* For a given value of θ $(-\pi<\theta<\pi)$, $\xi(t)$ is a single-valued function of t defined everywhere in the t-plane except when $\theta=0$ and t lies in the shaded region to the left of ER+EQ in figure 2(a). When $t>1$, $\xi(t)$ is positive provided that $|\theta|<\frac{3}{4}\pi$; this result is false if $\frac{3}{4}\pi<|\theta|<\pi$ for the cut AQ passes across the real axis as θ passes through $\frac{3}{4}\pi$.

Let $x\equiv x(t)$ now denote the branch $e^{2i\theta}\xi(t)$. We then take the x-domain $\mathbf{D}(\theta)$ of [9], section 2 to be the map of the function $x(t)$ from which circles surrounding the points A and E and of radius $\frac{1}{2}\delta$ have been removed, δ being a fixed small positive number.

The map of $x(t)$ has two sheets when $\theta=0$ and three sheets when $0<|\theta|<\pi$. It is illustrated in figure 3 for $\theta=0$ and $0<\theta<\frac{1}{2}\pi$. The passage from one sheet to another takes place across the positive real axis and, when $0<\theta<\frac{1}{2}\pi$, the join AE. The heavy lines correspond to the cuts in the t-plane and are boundaries of $\mathbf{D}(\theta)$.

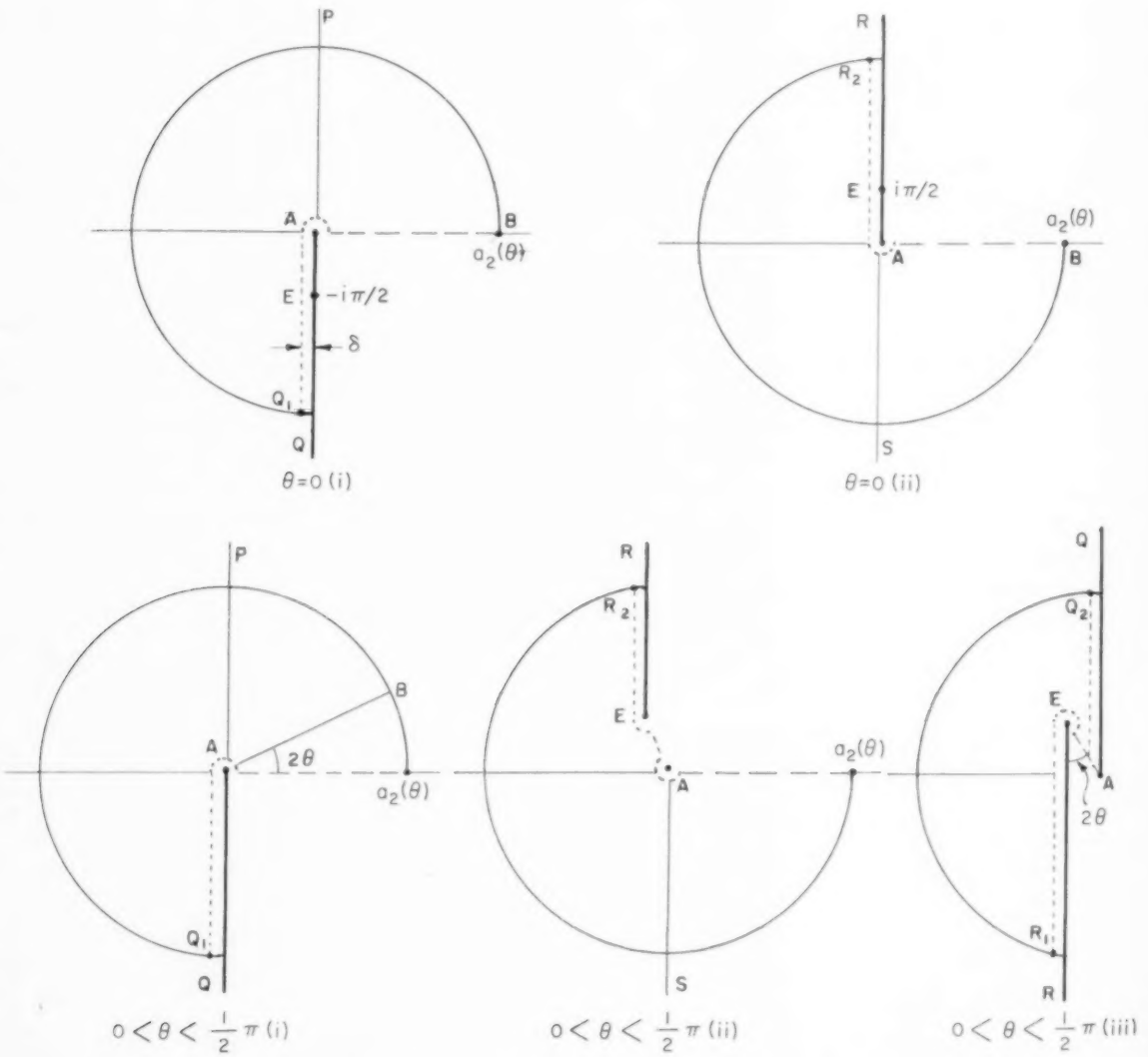


FIGURE 3. *x-plane: domains* $\mathbf{D}(\theta)$ *and* $\mathbf{G}(\theta)$.

The map of $x(t)$ is clearly a single unbounded domain. In order to ensure that $\mathbf{D}(\theta)$ is a single domain we have to impose the restriction $|\theta| \leq \pi - \epsilon$, omit sheet (iii) when $|\theta| < \epsilon$ and make $\delta = \delta(\epsilon)$ small enough. Otherwise, for example, the circle center E radius $\frac{1}{2}\delta$ would intersect the boundary AQ.

Next, we define the x-domain $\mathbf{G}(\theta)$ required by [9], section 2 to be that part of the map of $x(t)$ which remains after removing all points at a distance less than δ from its boundaries. Sheet (iii) is again to be omitted when $|\theta| < \epsilon$. The boundaries of $\mathbf{G}(\theta)$ are the dotted lines and circular arcs of figure 3; their t-maps are the dotted curves of figure 2. Clearly when $|\theta| \leq \pi - \epsilon$, $\mathbf{G}(\theta)$ is a single domain if $\delta = \delta(\epsilon)$ is small enough. We readily show that conditions (i) and (ii) of [9], section 2 are fulfilled (cf. [9], sec. 3(i)).

Following [9], section 2 further, we take $a_2(\theta)$ to be the point at infinity on the positive real axis. We then find that $\mathbf{H}_2(\theta) \equiv \mathbf{G}(\theta)$. From theorem A it now follows that there exists a solution W_2 of eq (3.8) such that for large positive values of u

$$W_2 \sim e^{-uz} \sum_{s=0}^{\infty} (-)^s \frac{\mathscr{A}_s^*(\theta, x)}{u^s} \qquad (x \in \mathbf{G}(\theta), \quad |\theta| \leq \pi - \epsilon), \tag{3.15}$$

uniformly with respect to x and θ. The coefficients are given by the recurrence relation

$$\mathscr{A}_{s+1}^*(\theta, x) = -\frac{1}{2}\frac{d}{dx}\mathscr{A}_s^*(\theta, x) + \frac{1}{2}\int e^{-4i\theta}\mathscr{F}(e^{-2i\theta}x)\mathscr{A}_s^*(\theta, x)dx, \tag{3.16}$$

the path of integration lying in $\mathbf{D}(\theta)$.

If we now restore the variables μ and ξ by means of eqs (3.7), and let

$$\mathscr{A}_s(\xi) \equiv (-)^s e^{2si\theta}\mathscr{A}_s^*(\theta, x), \tag{3.17}$$

the relations (3.15) and (3.16) become

$$W_2 \sim e^{-\mu^2\xi} \sum_{s=0}^{\infty} \frac{\mathscr{A}_s(\xi)}{\mu^{2s}} \qquad (t \in \mathbf{S}(\arg\mu), \quad |\arg\mu| \leq \pi - \epsilon) \tag{3.18}$$

as $|\mu| \to \infty$, where

$$\mathscr{A}_{s+1}(\xi) = \frac{1}{2}\mathscr{A}_s'(\xi) - \frac{1}{2}\int \mathscr{F}(\xi)\mathscr{A}_s(\xi)d\xi. \tag{3.19}$$

Here $\mathbf{S}(\arg\mu) \equiv \mathbf{S}(\theta)$ denotes the t-map of $\mathbf{G}(\theta)$; its boundaries are the dotted curves of figure 2.

It is of some interest to determine the asymptotic behavior of the boundaries of $\mathbf{S}(\theta)$. Consider, for example, the dotted curve AQ_1. Its equation may be written

$$\xi = (y - i\delta)e^{(\frac{3}{2}\pi - 2\theta)i}, \tag{3.20}$$

where the real parameter y ranges from 0 to $+\infty$. Reversion of (3.5) yields

$$t^2 = 2\xi + \frac{1}{2}\ln 8\xi + \frac{1}{2} + O(\xi^{-1}\ln\xi) \qquad \text{as } |\xi| \to \infty. \tag{3.21}$$

When y is large, we may substitute (3.20) in (3.21) and thence obtain

$$\arg t = \frac{3}{4}\pi - \theta + \frac{\cos 2\theta}{8}\frac{\ln 8y}{y} + \frac{\cos 2\theta + (2\theta - \frac{3}{2}\pi)\sin 2\theta - 4\delta}{8y} + O\left\{\left(\frac{\ln y}{y}\right)^2\right\}. \tag{3.22}$$

The corresponding equation of the principal curve AQ is of course obtained on setting $\delta = 0$ in this result. Clearly the principal curve AQ and the dotted curve AQ_1 ultimately always lie on the same side of their common asymptote, except when $\theta = \frac{3}{4}\pi$, in which event AQ is its own asymptote.

4. Identification of Solutions

The substitutions

$$a=-\frac{1}{2}\mu^2, \qquad z=\mu t\sqrt{2}, \tag{4.1}$$

transform eq (2.1) into (1.1). This result and the second of eqs (3.1) show that

$$W=(t^2-1)^{\frac{1}{4}}U\left(-\frac{1}{2}\mu^2,\mu t\sqrt{2}\right) \tag{4.2}$$

is a solution of (3.2). Equation (2.2) shows that when μ is fixed and $t\to\infty e^{-i\arg\mu}$ this solution becomes exponentially small. The same is true of the solution W_2 introduced in section 3, provided that $|\mu|$ is sufficiently large and $|\arg\mu|\leq\pi-\epsilon$; this can be seen from (3.18) and (3.5). Since all solutions which are linearly independent of (4.2) are exponentially large in the same circumstances, we see immediately that the solution (4.2) is a multiple of W_2.

Hence we have

$$U\left(-\frac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim g(\mu)\,\frac{e^{-\mu^2\xi}}{(t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}\frac{\mathscr{A}_s(\xi)}{\mu^{2s}} \qquad (t\,\epsilon\,\mathbf{S}(\arg\mu),\ \ |\arg\mu|\leq\pi-\epsilon) \tag{4.3}$$

as $|\mu|\to\infty$, uniformly with respect to t and $\arg\mu$. The branch of $(t^2-1)^{\frac{1}{4}}$ here and elsewhere is understood to be continuous in $\mathbf{S}(\arg\mu)$ and to be asymptotic to the principal value of $\sqrt{t}$ as $t\to\infty e^{-i\arg\mu}$. Similarly for the branch of $(t^2-1)^{\frac{1}{2}}$ used later. The function $g(\mu)$ is independent of t. Its asymptotic expansion for large $|\mu|$ can be found from the limiting form of the ratio of the two sides of the relation (4.3) as $t\to\infty e^{-i\arg\mu}$. Thus with the aid of (2.2) and (3.5) we find that

$$\frac{1}{g(\mu)}\sim 2^{\frac{1}{4}\mu^2+\frac{1}{4}}e^{\frac{1}{4}\mu^2}\mu^{-\frac{1}{2}\mu^2+\frac{1}{2}}\sum_{s=0}^{\infty}\frac{g_s}{\mu^{2s}} \qquad (|\arg\mu|\leq\pi-\epsilon) \tag{4.4}$$

as $|\mu|\to\infty$, uniformly with respect to $\arg\mu$. Here

$$g_s=\lim\,\mathscr{A}_s(\xi) \qquad \text{as } \xi\to\infty e^{-2i\arg\mu}. \tag{4.5}$$

That this limit exists and is independent of $\arg\mu$ can be seen from the lemma of [9], section 5. The g_s are not determined however until the arbitrary constants in the recurrence relation (3.19) have been fixed, and we now consider how to do this in the most advantageous manner.

From (3.19), (3.3) and the first of (3.1), we obtain

$$\mathscr{A}_{s+1}=\frac{1}{2}\,\frac{1}{(t^2-1)^{\frac{1}{2}}}\,\frac{d\mathscr{A}_s}{dt}+\frac{1}{8}\int\frac{3t^2+2}{(t^2-1)^{\frac{1}{2}}}\,\mathscr{A}_s\,dt. \tag{4.6}$$

Taking $\mathscr{A}_0=1$, we find

$$\mathscr{A}_1=\frac{t^3-6t}{24(t^2-1)^{\frac{1}{2}}}+\text{constant}. \tag{4.7}$$

A convenient value for this constant is zero, then $\mathscr{A}_1$ has the form of a polynomial in t divided by $(t^2-1)^{\frac{3}{2}}$.

The eq (4.7) suggests the substitution

$$\mathscr{A}_s=u_s(t)/(t^2-1)^{\frac{3}{2}s} \tag{4.8}$$

in (4.6). This leads to the recurrence relation

$$(t^2-1)u_s'(t)-3stu_s(t)=r_{s-1}(t), \tag{4.9}$$

where

$$8r_s(t)=(3t^2+2)u_s(t)-12(s+1)tr_{s-1}(t)+4(t^2-1)r_{s-1}'(t). \tag{4.10}$$

We can satisfy this relation by taking $u_s(t)$ to be a polynomial in t of degree $3s$ which is an odd or even function of t according as s is odd or even. All the coefficients are determined automatically if s is odd, but there is a degree of freedom if s is even, since the left of eq (4.9) vanishes when u_s is a multiple of $(t^2-1)^{\frac{3}{2}s}$. We remove this by making the coefficient of the highest power of t vanish.

Thus we prescribe that $u_s(t)$ is a polynomial in t of degree $3s$ (s odd), $3s-2$ (s even, $s\geq 2$). With $u_0(t)=1$ this condition determines the coefficients completely. Numerical calculation based on (4.9) and (4.10) yields

$$\left.\begin{aligned} &u_0(t)=1, \qquad u_1(t)=(t^3-6t)/24, \qquad u_2(t)=(-9t^4+249t^2+145)/1152,\\ &u_3(t)=(-4042t^9+18189t^7-28287t^5-1\ 51995t^3-2\ 59290t)/4\ 14720. \end{aligned}\right\} \tag{4.11}$$

Hence

$$g_{2s}=0 \quad (s=1,2,\ldots), \qquad g_0=1, \quad g_1=\frac{1}{24}, \quad g_3=-\frac{2021}{2\ 07360}. \tag{4.12}$$

An independent method of calculating the asymptotic expansion of $g(\mu)$ for large μ will be given in section 5.

Expansion for the derivative. Theorem A of [9] shows that the asymptotic series (4.3) may be differentiated term by term with respect to t. With the aid of the first of eqs (3.1) we deduce that

$$U'\left(-\frac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim -\frac{\mu}{\sqrt{2}}\,g(\mu)(t^2-1)^{\frac{1}{4}}e^{-\mu^2\xi}\sum_{s=0}^{\infty}\frac{\mathscr{B}_s(\xi)}{\mu^{2s}} \qquad (t\ \epsilon\ \mathbf{S}(\arg\mu), \quad |\arg\mu|\leq\pi-\epsilon) \tag{4.13}$$

as $|\mu|\to\infty$, uniformly with respect to t and $\arg\mu$. Here, in the notation of Miller,

$$U'(a,z)\equiv\partial U(a,z)/\partial z.$$

The coefficients are given by

$$\mathscr{B}_s(\xi)=v_s(t)/(t^2-1)^{\frac{3}{2}s}, \tag{4.14}$$

where

$$v_s(t)=u_s(t)+\frac{1}{2}\,tu_{s-1}(t)-r_{s-2}(t). \tag{4.15}$$

The first four $v_s(t)$ are

$$\left.\begin{aligned} &v_0(t)=1, \quad v_1(t)=(t^3+6t)/24, \quad v_2(t)=(15t^4-327t^2-143)/1152,\\ &v_3(t)=(-4042t^9+18189t^7-36387t^5+2\ 38425t^3+2\ 59290t)/4\ 14720. \end{aligned}\right\} \tag{4.16}$$

5. Use of Connection Formulas

In this section we seek expansions which hold in the complement of the domain $\mathbf{S}(\arg\mu)$. Since $U(a,z)$ is an integral function of a it will clearly suffice to obtain a set of expansions which, together with (4.3), cover the whole t-plane (except of course the neighborhoods of the points $t=\pm 1$) when $|\arg\mu|\leq\frac{1}{2}\pi$.

For convenience we introduce the notation $\mathbf{U}_j\equiv\mathbf{U}_j(\theta)$ ($j=0,\ 1,\ 2,\ 3,\ 4$) for the closed domains into which the t-plane is divided by the principal curves defined in section 3. Figure 4 shows the enumeration of these regions when $0\leq\theta<\pi$. When $-\pi<\theta\leq 0$ we define

$$\mathbf{U}_j(\theta)=\mathbf{U}_j^*(-\theta) \quad (j=0,2,4), \qquad \mathbf{U}_1(\theta)=\mathbf{U}_3^*(-\theta), \quad \mathbf{U}_3(\theta)=\mathbf{U}_1^*(-\theta), \tag{5.1}$$

where the star denotes the conjugate region. Then the boundaries of each region $\mathbf{U}_j(\theta)$ vary continuously with θ in the interval $(-\pi,\pi)$.

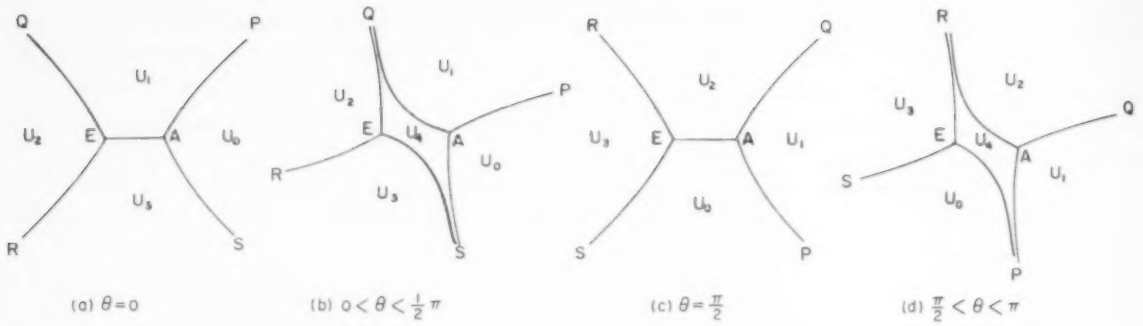


FIGURE 4. *t-plane: domains* $\mathbf{U}_j(\theta)$.

The following relations, suggested by figure 4, are easily verified:

$$\left.\begin{aligned} &\mathbf{U}_0(\theta+\tfrac{1}{2}\pi)=\mathbf{U}_3(\theta),\quad \mathbf{U}_1(\theta+\tfrac{1}{2}\pi)=\mathbf{U}_0(\theta),\quad \mathbf{U}_2(\theta+\tfrac{1}{2}\pi)=\mathbf{U}_1(\theta),\\ &\mathbf{U}_3(\theta+\tfrac{1}{2}\pi)=\mathbf{U}_2(\theta),\quad \mathbf{U}_4(\theta+\tfrac{1}{2}\pi)=\mathbf{U}_4(\theta),\end{aligned}\right\} \tag{5.2}$$

$$\left.\begin{aligned} &\mathbf{U}_0(\theta+\pi)=\mathbf{U}_2(\theta),\quad \mathbf{U}_1(\theta+\pi)=\mathbf{U}_3(\theta),\quad \mathbf{U}_2(\theta+\pi)=\mathbf{U}_0(\theta),\\ &\mathbf{U}_3(\theta+\pi)=\mathbf{U}_1(\theta),\quad \mathbf{U}_4(\theta+\pi)=\mathbf{U}_4(\theta).\end{aligned}\right\} \tag{5.3}$$

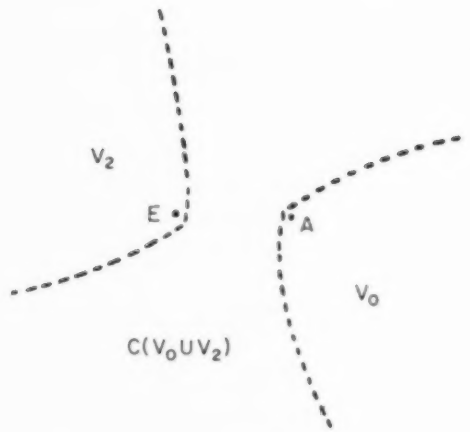


FIGURE 5. *t-plane* $\left(0<\theta<\frac{1}{2}\pi\right)$.

The region formed by adding to $\mathbf{U}_j(\theta)$ points whose x-maps (cf. fig. 3) are at a distance less than δ from the x-map of the boundary of $\mathbf{U}_j(\theta)$ we shall denote by $\mathbf{V}_j(\theta)$ or $\mathbf{V}_j$. Figure 5 illustrates for example $\mathbf{V}_0(\theta)$, $\mathbf{V}_2(\theta)$, and $C\{\mathbf{V}_0(\theta)\cup\mathbf{V}_2(\theta)\}$, the complement of the union of $\mathbf{V}_0(\theta)$ and $\mathbf{V}_2(\theta)$, when $0<\theta<\frac{1}{2}\pi$. The boundaries are of the same character as the dotted curves of figure 2.

The first connection formula we use is

$$U\left(-\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)=(2\pi)^{-\frac{1}{2}}\Gamma\left(\tfrac{1}{2}+\tfrac{1}{2}\mu^2\right)\left\{e^{i\pi(-\frac{1}{4}\mu^2+\frac{1}{4})}U\left(\tfrac{1}{2}\mu^2,-i\mu t\sqrt{2}\right)+e^{i\pi(\frac{1}{4}\mu^2-\frac{1}{4})}U\left(\tfrac{1}{2}\mu^2,i\mu t\sqrt{2}\right)\right\}, \tag{5.4}$$

obtained from (2.5). Replacing μ by $\mp i\mu$ in (4.3), we have

$$U\left(\tfrac{1}{2}\mu^2,-i\mu t\sqrt{2}\right)\sim g(-i\mu)\,\frac{e^{\mu^2\xi}}{(t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}(-)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\qquad \left(t\,\epsilon\,\mathbf{S}\left(\theta-\tfrac{1}{2}\pi\right),\quad -\tfrac{1}{2}\pi+\epsilon\leq\theta\leq\tfrac{3}{2}\pi-\epsilon\right), \tag{5.5}$$

$$U\left(\frac{1}{2}\mu^2, i\mu t\sqrt{2}\right)\sim g(i\mu)\,\frac{e^{\mu^2\xi}}{(t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}(-)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\qquad\left(t\,\epsilon\,\mathbf{S}\left(\theta+\frac{1}{2}\pi\right),\quad -\frac{3}{2}\pi+\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon\right),\tag{5.6}$$

and from (4.4) and (4.12),

$$g(\mp i\mu)\sim e^{\pm i\pi(\frac{1}{4}\mu^2+\frac{1}{4})}2^{\frac{1}{4}\mu^2-\frac{1}{4}}e^{\frac{1}{4}\mu^2}\mu^{-\frac{1}{2}\mu^2-\frac{1}{2}}\left(1-\sum_{s=0}^{\infty}\frac{g_{2s+1}}{\mu^{4s+2}}\right)^{-1}\qquad\left\{\begin{matrix}-\frac{1}{2}\pi+\epsilon\leq\theta\leq\frac{3}{2}\pi-\epsilon\\ -\frac{3}{2}\pi+\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon\end{matrix}\right\}.\tag{5.7}$$

When $|\theta|\leq\frac{1}{2}\pi-\epsilon$, we can substitute (5.5) and (5.6) in (5.4). Care is required, however, because at any given point in the t-plane the branches of $\xi=\xi(t)$ and $(t^2-1)^{\frac{1}{4}}$ may not be the same in the two formulas.

When $0\leq\theta\leq\frac{1}{2}\pi-\epsilon$, $\mathbf{S}(\theta-\frac{1}{2}\pi)$ is the complement of the region enclosed by "dotted curves" surrounding the principal curves AS and ER of figures 2(a) and 2(b); $\mathbf{S}(\theta+\frac{1}{2}\pi)$ is the complement of the region enclosed by dotted curves surrounding AP and EQ. To the right of AP+AS, i.e., in $\mathbf{U}_0(\theta)$, the branches of $\xi(t)$ and $(t^2-1)^{\frac{1}{4}}$ are the same in (5.5) and (5.6). Substitution of these formulas in (5.4) leads to a nugatory result, however, as we would expect, since $e^{\mu^2\xi}$ is exponentially large here whereas $U(-\frac{1}{2}\mu^2,\ \mu t\sqrt{2})$ is exponentially small.

In the central region between AP+AS and EQ+ER, i.e., in $C(\mathbf{U}_0\cup\mathbf{U}_2)$, the two branches of $\xi(t)$ have opposite signs. Let

$$\xi=-i\eta\text{ in (5.5)},\qquad \xi=i\eta\text{ in (5.6)}.\tag{5.8}$$

Then η is real and positive when $-1<t<1$; from (3.4) we find

$$\eta=\int_t^1(1-t^2)^{\frac{1}{2}}dt=\frac{1}{2}\cos^{-1}t-\frac{1}{2}\,t(1-t^2)^{\frac{1}{2}}.\tag{5.9}$$

The appropriate branches of $(t^2-1)^{\frac{1}{4}}$ and $\mathscr{A}_s(\xi)$ in (5.5) are given by

$$(t^2-1)^{\frac{1}{4}}=e^{\frac{1}{4}i\pi}(1-t^2)^{\frac{1}{4}},\qquad \mathscr{A}_s(\xi)=e^{-\frac{3}{2}is\pi}u_s(t)/(1-t^2)^{\frac{3}{2}s}\tag{5.10}$$

(cf. (4.8)), in which $(1-t^2)^{\frac{1}{4}}$ and $(1-t^2)^{\frac{3}{2}s}$ have their principal values when $-1<t<1$. For (5.6) we change the sign of i.

Carrying out the substitution, we obtain

$$U(-\tfrac{1}{2}\mu^2,\ \mu t\sqrt{2})\sim\frac{2\tilde{g}(\mu)}{(1-t^2)^{\frac{1}{4}}}\left[\cos\,(\mu^2\eta-\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s}(t)}{(1-t^2)^{3s}\mu^{4s}}-\sin\,(\mu^2\eta-\tfrac{1}{4}\,\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s+1}(t)}{(1-t^2)^{3s+\frac{3}{2}}\mu^{4s+2}}\right]\tag{5.11}$$

as $|\mu|\to\infty$, uniformly with respect to $t\,\epsilon\,C(\mathbf{V}_0\cup\mathbf{V}_2)$ and θ in the interval $(-\frac{1}{2}\pi+\epsilon,\ \frac{1}{2}\pi-\epsilon)$. Here

$$\begin{aligned}\tilde{g}(\mu)&\sim(2\pi)^{-\frac{1}{2}}\Gamma(\tfrac{1}{2}+\tfrac{1}{2}\mu^2)2^{\frac{1}{4}\mu^2-\frac{1}{4}}e^{\frac{1}{4}\mu^2}\mu^{-\frac{1}{2}\mu^2-\frac{1}{2}}\left\{1-\sum_{s=0}^{\infty}\frac{g_{2s+1}}{\mu^{4s+2}}\right\}_{-1}\\ &\sim 2^{-\frac{1}{4}\mu^2-\frac{1}{4}}e^{-\frac{1}{4}\mu^2}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}\left\{\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\right\}\left\{1-\sum_{s=0}^{\infty}\frac{g_{2s+1}}{\mu^{4s+2}}\right\}_{-1}\qquad(|\theta|\leq\tfrac{1}{2}\pi-\epsilon),\end{aligned}\tag{5.12}$$

on substituting for $\Gamma(\frac{1}{2}+\frac{1}{2}\mu^2)$ by means of (2.22).

The expansion (5.11) is one of the required formulas. Before leaving it we shall show that $\tilde{g}(\mu)=g(\mu)$, and in the process obtain a new derivation of the coefficients in the asymptotic expansion (4.4) for $1/g(\mu)$.

When $\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon$ the principal curve AQ is included in the region of validity of (5.11); this is easily proved by considering the x-map. On this curve $\mu^2\eta$ is positive. Let us set

$$\eta=\eta_m\equiv\left(m+\frac{1}{4}\right)\frac{\pi}{\mu^2},\tag{5.13}$$

where m is a positive integer. Then (5.11) reduces to

$$U\left(-\frac{1}{2}\mu^2,\ \mu t_m\sqrt{2}\right)\sim(-)^m\frac{2\tilde{g}(\mu)}{(1-t_m^2)^{\frac{1}{4}}}\sum_{s=0}^{\infty}\frac{(-)^s u_{2s}(t_m)}{(1-t_m^2)^{3s}\mu^{4s}}, \tag{5.14}$$

where t_m is the value of t corresponding to $\eta=\eta_m$.

Now keep μ fixed and let $m\to\infty$. Then $t_m\to\infty e^{i(\frac{3}{4}\pi-\theta)}$, and so from (2.7)

$$U\left(-\frac{1}{2}\mu^2,\ \mu t_m\sqrt{2}\right)\sim e^{-\frac{1}{2}\mu^2 t_m^2}(\mu t_m\sqrt{2})^{\frac{1}{2}\mu^2-\frac{1}{2}}+\frac{(2\pi)^{\frac{1}{2}}e^{i\pi(\frac{1}{2}\mu^2+\frac{1}{2})}}{\Gamma(\frac{1}{2}-\frac{1}{2}\mu^2)}\,e^{\frac{1}{2}\mu^2 t_m^2}(\mu t_m\sqrt{2})^{-\frac{1}{2}\mu^2-\frac{1}{2}}. \tag{5.15}$$

From (3.5) and the first of (5.8)

$$\frac{1}{2}t_m^2-\frac{1}{2}\ln 2t_m=-i\eta_m+\frac{1}{4}+O(|t_m|^{-2}). \tag{5.16}$$

Hence

$$e^{-\frac{1}{2}\mu^2 t_m^2}(\mu t_m\sqrt{2})^{\frac{1}{2}\mu^2-\frac{1}{2}}\sim(-)^m e^{\frac{1}{4}\pi i-\frac{1}{4}\mu^2}2^{-\frac{1}{4}\mu^2-\frac{1}{4}}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}t_m^{-\frac{1}{2}},$$

$$e^{\frac{1}{2}\mu^2 t_m^2}(\mu t_m\sqrt{2})^{-\frac{1}{2}\mu^2-\frac{1}{2}}\sim(-)^m e^{-\frac{1}{4}\pi i+\frac{1}{4}\mu^2}2^{\frac{1}{4}\mu^2-\frac{1}{4}}\mu^{-\frac{1}{2}\mu^2-\frac{1}{2}}t_m^{-\frac{1}{2}},$$

and substituting these results in (5.15), we obtain

$$(-)^m e^{\frac{1}{4}\pi i-\frac{1}{4}\mu^2}2^{-\frac{1}{4}\mu^2-\frac{1}{4}}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}\left\{1+\frac{(2\pi)^{\frac{1}{2}}e^{\frac{1}{2}i\pi\mu^2+\frac{1}{2}\mu^2}2^{\frac{1}{2}\mu^2}\mu^{-\mu^2}}{\Gamma(\frac{1}{2}-\frac{1}{2}\mu^2)}\right\}t_m^{-\frac{1}{2}}, \tag{5.17}$$

as the limiting form of the left-hand side of (5.14).

On the right of (5.14) we have

$$(1-t_m^2)^{\frac{1}{4}}\sim e^{-\frac{1}{4}\pi i}t_m^{\frac{1}{2}},\qquad u_{2s}(t_m)/(1-t_m^2)^{3s}\to 0\quad(s\geq 1)$$

(cf. (5.10), (4.5) and (4.12)). Comparing the two sides we see that

$$\tilde{g}(\mu)\simeq 2^{-\frac{1}{4}\mu^2-\frac{1}{4}}e^{-\frac{1}{4}\mu^2}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}\left\{\frac{1}{2}+\frac{(2\pi)^{\frac{1}{2}}e^{\frac{1}{2}i\pi\mu^2+\frac{1}{2}\mu^2}2^{\frac{1}{2}\mu^2}\mu^{-\mu^2}}{2\Gamma(\frac{1}{2}-\frac{1}{2}\mu^2)}\right\}\qquad(\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon), \tag{5.18}$$

where we use the symbol $\simeq$ to mean that the two sides of this relation have the same asymptotic expansion for large $|\mu|$.

We now substitute (2.27), with $z=-\frac{1}{2}\mu^2$, in (5.18) and obtain

$$\tilde{g}(\mu)\sim 2^{-\frac{1}{4}\mu^2-\frac{1}{4}}e^{-\frac{1}{4}\mu^2}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}\left\{\frac{1}{2}+\frac{1}{2}\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\right\}. \tag{5.19}$$

Although this result has been proved on the assumption that $\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon$, comparison with (5.12) shows immediately that it holds in the range $|\theta|\leq\frac{1}{2}\pi-\epsilon$, and that

$$\frac{1}{2}+\frac{1}{2}\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\sim\left\{\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\right\}\left\{1-\sum_{s=0}^{\infty}\frac{g_{2s+1}}{\mu^{4s+2}}\right\}^{-1}, \tag{5.20}$$

identically. Thence using (2.26) and changing the sign of μ^2, we find

$$\left\{1+\sum_{s=0}^{\infty}\frac{g_{2s+1}}{\mu^{4s+2}}\right\}\left\{\frac{1}{2}+\frac{1}{2}\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\right\}\sim 1. \tag{5.21}$$

Multiplying this identity with (5.19), and comparing the result with (4.4) (remembering that $g_{2s}=0$ if $s\geq 1$), we see that we can take

$$\tilde{g}(\mu)=g(\mu) \tag{5.22}$$

in the expansion (5.11), as stated earlier.

We can express the coefficients g_{2s+1} in terms of γ_s, by equating coefficients in (5.21); thus

$$g_1=-\gamma_1=\frac{1}{24}, \qquad g_3=-2g_1\gamma_2-4\gamma_3=-\frac{2021}{2\,07360},$$

on using (2.25). These values agree with those given in (4.12).

The expansion (5.11) is the first of our expansions for $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ in regions complementary to $\mathbf{S}(\theta)$. The corresponding expansion for the derivative is

$$U'(-\tfrac{1}{2}\mu^2, \mu t\sqrt{2}) \sim \mu\sqrt{2}g(\mu)(1-t^2)^{\frac{1}{4}}\left[\sin(\mu^2\eta-\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s v_{2s}(t)}{(1-t^2)^{3s}\mu^{4s}}\right.$$
$$\left.+\cos(\mu^2\eta-\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s v_{2s+1}(t)}{(1-t^2)^{3s+\frac{3}{2}}\mu^{4s+2}}\right], \quad (5.23)$$

valid under the same conditions as (5.11); the coefficients $v_s(t)$ are given by (4.15) and (4.16).

A further expansion can be found by substituting (5.5) and (5.6) in (5.4) when t lies in the region ($\mathbf{U}_2$) to the left of EQ+ER. We shall not carry out this substitution however, because the expansion so obtained is contained in a more general result (5.28), which we shall now derive. As a preliminary we record the following relations in which the symbol $\simeq$ has the same meaning as in (5.18).

$$g(\mu e^{i\pi}) \simeq e^{i\pi(\frac{1}{2}\mu^2-\frac{1}{2})}g(\mu) \qquad (-\pi+\epsilon\leq\arg\mu\leq-\epsilon), \quad (5.24)$$

$$\left.\begin{aligned}\frac{(2\pi)^{\frac{1}{2}}e^{-i\pi(\frac{1}{4}\mu^2+\frac{1}{4})}}{\Gamma(\frac{1}{2}-\frac{1}{2}\mu^2)}g(\mu e^{-\frac{1}{2}i\pi}) &\simeq 2\cos\frac{1}{2}\pi\mu^2\cdot g(\mu) \qquad (|\arg\mu|\leq\frac{1}{2}\pi-\epsilon),\\ &\simeq e^{-\frac{1}{2}i\pi\mu^2}g(\mu) \qquad (\epsilon\leq\arg\mu\leq\pi-\epsilon).\end{aligned}\right\} \quad (5.25)$$

We use the connection formula

$$U\left(-\frac{1}{2}\mu^2,-\mu t\sqrt{2}\right)=e^{i\pi(-\frac{1}{2}\mu^2+\frac{1}{2})}U\left(-\frac{1}{2}\mu^2,\mu t\sqrt{2}\right)+\frac{(2\pi)^{\frac{1}{2}}e^{-i\pi(\frac{1}{4}\mu^2+\frac{1}{4})}}{\Gamma(\frac{1}{2}-\frac{1}{2}\mu^2)}U\left(\frac{1}{2}\mu^2,-i\mu t\sqrt{2}\right), \quad (5.26)$$

obtained from (2.4). We may substitute the expansions (4.3) and (5.5) when

$$-\tfrac{1}{2}\pi+\epsilon\ \leq\theta\leq\pi-\epsilon.$$

The branches of $\xi(t)$ are the same in both expansions when

$$\left.\begin{aligned}&t\in C(\mathbf{V}_2\cup\mathbf{V}_3\cup\mathbf{V}_4) \qquad \left(0\leq\theta\leq\frac{1}{2}\pi\right),\\ &t\in C(\mathbf{V}_2\cup\mathbf{V}_3) \qquad \left(-\frac{1}{2}\pi+\epsilon\leq\theta\leq 0,\ \ \frac{1}{2}\pi\leq\theta\leq\pi-\epsilon\right).\end{aligned}\right\} \quad (5.27)$$

Using (5.25) we obtain immediately

$$\left.\begin{aligned}U\left(-\frac{1}{2}\mu^2,-\mu t\sqrt{2}\right) \sim \frac{g(\mu)}{(t^2-1)^{\frac{1}{4}}}&\left[\left(\sin\frac{1}{2}\pi\mu^2+i\cos\frac{1}{2}\pi\mu^2\right)e^{-\mu^2\xi}\sum_{s=0}^{\infty}\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\right.\\ &\left.+2\cos\frac{1}{2}\pi\mu^2\cdot e^{\mu^2\xi}\sum_{s=0}^{\infty}(-)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\right]\\ &t\in C(\mathbf{V}_2\cup\mathbf{V}_3) \qquad \left(-\frac{1}{2}\pi+\epsilon\leq\theta\leq 0\right),\\ &t\in C(\mathbf{V}_2\cup\mathbf{V}_3\cup\mathbf{V}_4) \qquad \left(0\leq\theta\leq\frac{1}{2}\pi-\epsilon\right),\end{aligned}\right\} \quad (5.28)$$

and

$$U\left(-\frac{1}{2}\mu^2,-\mu t\sqrt{2}\right)\sim e^{-\frac{1}{2}i\pi\mu^2}\frac{g(\mu)}{(t^2-1)^{\frac{1}{4}}}\left[ie^{-\mu^2\xi}\sum_{s=0}^{\infty}\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}+e^{\mu^2\xi}\sum_{s=0}^{\infty}(-)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\right]$$

$$t\in C(\mathbf{V}_2\cup\mathbf{V}_3\cup\mathbf{V}_4)\qquad\left(\epsilon\leq\theta\leq\frac{1}{2}\pi\right),$$

$$t\in C(\mathbf{V}_2\cup\mathbf{V}_3)\qquad\left(\frac{1}{2}\pi\leq\theta\leq\pi-\epsilon\right). \tag{5.29}$$

The last result may be expressed in a form more pertinent to the range $|\theta|\leq\frac{1}{2}\pi$ by replacing μ by $\mu e^{i\pi}$ and using (5.24). This gives

$$U\left(-\frac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim\frac{g(\mu)}{(t^2-1)^{\frac{1}{4}}}\left[e^{-\mu^2\xi}\sum_{s=0}^{\infty}\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}-ie^{\mu^2\xi}\sum_{s=0}^{\infty}(-)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\right]$$

$$t\in C(\mathbf{V}_0\cup\mathbf{V}_1)\qquad\left(-\frac{1}{2}\pi\leq\theta\leq-\epsilon\right). \tag{5.30}$$

Three further formulas, (5.28*), (5.29*), and (5.30*), can be derived immediately from (5.28), (5.29), and (5.30), respectively, merely by changing the sign of i. They are valid with the conditions

$$\left.\begin{aligned}&t\in C(\mathbf{V}_1\cup\mathbf{V}_2\cup\mathbf{V}_4)\qquad\left(-\frac{1}{2}\pi+\epsilon\leq\theta\leq 0\right),\\&t\in C(\mathbf{V}_1\cup\mathbf{V}_2)\qquad\left(0\leq\theta\leq\frac{1}{2}\pi-\epsilon\right),\end{aligned}\right\}\text{for (5.28*),}$$

$$\left.\begin{aligned}&t\in C(\mathbf{V}_1\cup\mathbf{V}_2)\qquad\left(-\pi+\epsilon\leq\theta\leq-\frac{1}{2}\pi\right),\\&t\in C(\mathbf{V}_1\cup\mathbf{V}_2\cup\mathbf{V}_4)\qquad\left(-\frac{1}{2}\pi\leq\theta\leq-\epsilon\right),\end{aligned}\right\}\text{for (5.29*),}$$

$$t\in C(\mathbf{V}_0\cup\mathbf{V}_3)\qquad\left(\epsilon\leq\theta\leq\frac{1}{2}\pi\right),\qquad\text{for (5.30*).}$$

6. Summary of Expansions in Terms of Elementary Functions

The principal expansion of this kind is (4.3). It is valid for large $|\mu|$ uniformly with respect to $\arg\mu\equiv\theta$ and t when $-\pi+\epsilon\leq\theta\leq\pi-\epsilon$ and t lies in the unbounded domain $\mathbf{S}(\theta)$, the boundaries of which are the dotted curves illustrated in figure 2. The function $\xi(t)$ is given by (3.4), that branch being chosen which is continuous in $\mathbf{S}(\theta)$ and takes the value (3.13) or (3.14) at the origin. The branch of $(t^2-1)^{\frac{1}{4}}$ in (4.3) is continuous in $\mathbf{S}(\theta)$ and is asymptotic to the principal value of $\sqrt{t}$ as $t\to\infty e^{-i\theta}$. Both $\xi(t)$ and $(t^2-1)^{\frac{1}{4}}$ are positive when $t>1$ and $|\theta|<\frac{3}{4}\pi$. The coefficients $\mathscr{A}_s(\xi)$ are given by (4.8) and (4.11), the branch of $(t^2-1)^{\frac{1}{2}s}$ being defined similarly. The function $g(\mu)$ is calculable from either of the asymptotic expansions

$$\frac{1}{g(\mu)}\sim 2^{\frac{1}{4}\mu^2+\frac{1}{4}}e^{\frac{1}{4}\mu^2}\mu^{-\frac{1}{2}\mu^2+\frac{1}{2}}\left\{1+\sum_{s=0}^{\infty}\frac{g_{2s+1}}{\mu^{4s+2}}\right\}, \tag{6.1}$$

$$g(\mu)\sim 2^{-\frac{1}{4}\mu^2-\frac{1}{4}}e^{-\frac{1}{4}\mu^2}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}\left\{1+\frac{1}{2}\sum_{s=1}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\right\}, \tag{6.2}$$

both of which are uniformly valid with respect to $\arg\mu$ in the interval $(-\pi+\epsilon,\ \pi-\epsilon)$. Here g_1, g_3 are given by (4.12); γ_1, γ_2, γ_3, γ_4 by (2.25).

Expansions valid in regions complementary to $\mathbf{S}(\theta)$ are (5.11), (5.28), (5.29), (5.30), (5.28*), (5.29*), and (5.30*). When $0\leq\theta\leq\frac{1}{2}\pi$ the expansions (5.28) and (5.29) between them cover the principal curve ER of figure 2; expansions (5.28*) and (5.30*) cover the principal curve AQ, except when $0\leq\theta\leq\epsilon$ and $\operatorname{Re} t\geq -1-\delta'$. The region of validity of (5.11) however includes that part of AQ to the right of $\operatorname{Re} t\geq -1+\delta'$ when $0\leq\theta\leq\epsilon$. Hence the expansions we have obtained cover the whole t-plane (except of course the neighborhoods of the points $t=\pm 1$) when $0\leq\theta\leq\frac{1}{2}\pi$, and similarly also when $-\frac{1}{2}\pi\leq\theta\leq 0$.

The branches of the functions ξ, $(t^2-1)^{\frac{1}{4}}$ and $\mathscr{A}_s(\xi)$ in (5.28), (5.29), (5.30), and the conjugate results are the same as in (4.3), and the function $g(\mu)$ again has the asymptotic expansions (6.1) and (6.2).

In (5.11) $\tilde{g}(\mu)$ is identical with $g(\mu)$, and $\eta=\eta(t)$ is given by (5.9). The branch of the latter function is continuous in the region of validity of (5.11) and positive when $-1<t<1$; the same is true of the branches of $(1-t^2)^{\frac{1}{4}}$ and $(1-t^2)^{\frac{3}{4}}$.

The expansions for the derivative $U'(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ corresponding to (4.3) and (5.11) are (4.13) and (5.23) respectively, the coefficients $\mathscr{B}_s(\xi)$ and $v_s(t)$ being given by (4.14) to (4.16). The derivative expansions corresponding to (5.28) to (5.30*) have not been recorded, but they can be deduced immediately by use of the formal identities

$$\frac{d}{dt}\,\frac{e^{\pm\mu^2\xi}}{(t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}(\mp)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}=\pm\mu^2(t^2-1)^{\frac{1}{4}}e^{\pm\mu^2\xi}\sum_{s=0}^{\infty}(\mp)^s\frac{\mathscr{B}_s(\xi)}{\mu^{2s}}, \tag{6.3}$$

and are valid under the same conditions as the parent formulas.

Uniform asymptotic expansions of solutions of eq (1.1) for large $|\mu|$ and unbounded t have previously been given by Watson [12], Schwid [10], and Darwin [1].

Watson applies the method of steepest descents to integral representations of

$$D_n(z)\equiv U\left(-n-\frac{1}{2}, z\right)$$

and obtains a set of asymptotic expansions for the function $D_n(2\zeta\sqrt{n})$ when $|n|$ is large and $|\arg n|\leq\frac{1}{2}\pi-\epsilon$, which cover the whole ζ-plane other than the neighborhoods of the points $\zeta=\pm 1$. For each expansion the ζ-region of validity is fixed, unlike the regions obtained in the present paper which are permitted to depend on $\arg n$. For example, the region of validity of Watson's expansion analogous to our principal expansion (4.3) is

$$\mathbf{U}_0\left(\frac{1}{4}\pi\right)\cup\mathbf{U}_1\left(\frac{1}{4}\pi\right)\quad(|\zeta-1|\geq\delta).$$

This is, as we would expect, the region common to $\mathbf{S}(\arg\mu)$ as $\arg\mu$ varies over the interval $(-\frac{1}{4}\pi+\epsilon, \frac{1}{4}\pi-\epsilon)$.

Schwid applies the methods of Langer for the asymptotic solution of differential equations and his results resemble those we have obtained. The main differences are that Schwid confines attention to the regions $\operatorname{Re} t\geq 0$ and $|\arg\mu|<\frac{3}{4}\pi$, gives only the leading terms in each expansion and identifies the asymptotic solutions by considering their behavior at $t=0$ rather than $t=\infty$.

Schwid's results apply to the even and odd solutions of the differential equation. An effect of carrying out the identification at $t=0$ is that the error term in the asymptotic representation of each of these solutions is obtained essentially in the form

$$e^{-\mu^2\xi}O(|\mu|^{-2})+e^{\mu^2\xi}O(|\mu|^{-2}) \tag{6.4}$$

in the present notation. If we try to deduce from these results the asymptotic form of the function $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ by means of connection formulas, the error term will necessarily also be of the form (6.4). Accordingly, a nugatory result will be obtained in the region $(\mathbf{U}_0(\arg\mu))$ in which $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ is exponentially small for large $|\mu|$. Indeed, the proof given on page

358 of Schwid's paper of the fact that the error term in the asymptotic expression for the Hermite polynomials is of the form $e^{-\mu^2\xi}O(|\mu|^{-2})$ instead of (6.4) is false. The mistake is apparently due to overlooking the fact that the O's are functions of the independent variable as well as of the parameter μ.

Darwin develops series solutions of the eq (2.10), and in [4] Miller applies the same methods to eq (2.1). The results are typified by the series

$$U(a,z)=\frac{\{\Gamma(\frac{1}{2}-a)\}^{\frac{1}{2}}}{(2\pi)^{\frac{1}{4}}}\exp\left\{-\theta-\frac{1}{2}\ln X+\sum_{s=1}^{\infty}(-)^s\frac{d_{3s}}{X^{3s}}\right\} \tag{6.5}$$

for negative a and real positive z, in which

$$X=\sqrt{(z^2+4a)}, \qquad \theta=\frac{1}{2}\int_{z\sqrt{(-a)}}^{z}X\,dz,$$

and d_{3s} is a polynomial in z of degree $3s$ if s is odd, or s if s is even. Explicit expressions for $d_3, d_6, \ldots, d_{24}$ are given on page 68 of [4].

The series (6.5) is derived by purely formal methods from the differential equation without any investigation of its asymptotic nature. Applying the substitutions $a=-\frac{1}{2}\mu^2$ and $z=\mu t\sqrt{2}$ we find that $X=\mu(t^2-1)^{\frac{1}{2}}\sqrt{2}$ and $\theta=\mu^2\xi(t)$, in the notation of sections 3 and 4. Next, replacing $\{\Gamma(\frac{1}{2}-a)\}^{\frac{1}{2}}\equiv\{\Gamma(\frac{1}{2}+\frac{1}{2}\mu^2)\}^{\frac{1}{2}}$ by its asymptotic expansion for large positive μ (cf. (2.22)) and expanding $\exp\{\sum(-)^s d_{3s}X^{-3s}\}$ in descending powers of X^3, we find that the series (6.5) reduces to the form (4.3). Hence (6.5) must be a uniform asymptotic expansion for large $|a|$ of the same character as (4.3).

Darwin remarks that the explicit expressions for the coefficients d_{3s} are simpler in form and easier to calculate than the expressions for the coefficients in series of the form (4.3). This is indeed true, but we cannot however regard the form (6.5) as being entirely superior to (4.3); it does not, for example, lead to convenient forms of expansion when we differentiate with respect to z.

Part 3. Expansions in Terms of Airy Functions

7. Asymptotic Solutions of the Differential Equation

In this section we seek to apply theorem B of [9] to the differential eq (1.1). The preliminary step is to take new variables $\zeta=\zeta(t)$ and W, defined by

$$\zeta\left(\frac{d\zeta}{dt}\right)^2=t^2-1, \qquad W=\left(\frac{d\zeta}{dt}\right)^{\frac{1}{2}}w=\left(\frac{t^2-1}{\zeta}\right)^{\frac{1}{4}}w \tag{7.1}$$

(cf. [7], sec. 2). Equation (1.1) then becomes

$$\frac{d^2W}{d\zeta^2}=\{\mu^4\zeta+F(\zeta)\}W, \tag{7.2}$$

where

$$F(\zeta)\equiv\left(\frac{dt}{d\zeta}\right)^{\frac{1}{2}}\frac{d^2}{d\zeta^2}\left\{\left(\frac{dt}{d\zeta}\right)^{-\frac{1}{2}}\right\}=\frac{5}{16\zeta^2}+\zeta\mathscr{F}(\xi)=\frac{5}{16\zeta^2}-\frac{(3t^2+2)}{4(t^2-1)^3}\zeta \tag{7.3}$$

on reduction (cf. (3.3)).

The first of eqs (7.1) may be integrated to give

$$\frac{2}{3}\zeta^{\frac{3}{2}}=\int_1^t(t^2-1)^{\frac{1}{2}}dt=\frac{1}{2}t(t^2-1)^{\frac{1}{2}}-\frac{1}{2}\ln\{t+(t^2-1)^{\frac{1}{2}}\}, \tag{7.4}$$

on specifying that $\zeta=0$ when $t=1$. In terms of the variable ξ of section 3,

$$\frac{2}{3}\zeta^{\frac{3}{2}}=\xi. \tag{7.5}$$

An essential difference between the functions $\xi(t)$ and $\zeta(t)$ is that the former has a branch point at $t=1$, whereas the latter is analytic at this point. Both functions have branch points at $t=-1$. The branch of $\zeta(t)$ which will be the most convenient for our purposes depends on $\theta \equiv \arg \mu$. We define it to be positive when $t>1$ and continuous in the t-plane cut along the principal curve ER of figure 2 when $0<\theta<\pi$, along the curve conjugate to ER when $-\pi<\theta<0$, or along EQ and ER when $\theta=0$. This determines $\zeta(t)$ uniquely everywhere except in the region $(\mathbf{U}_2(0))$ to the left of EQ+ER when $\theta=0$.

Following [9], section 1, we make the further substitutions

$$u=|\mu|^2, \qquad \theta=\arg \mu, \qquad \zeta=e^{-\frac{4}{3}i\theta}z. \tag{7.6}$$

Equation (7.2) then becomes

$$\frac{d^2W}{dz^2}=\{u^2z+e^{-\frac{5}{3}i\theta}F(e^{-\frac{4}{3}i\theta}z)\}W. \tag{7.7}$$

The mapping of the t-plane on the z-plane is conveniently carried out by using $x(t)$ (defined in sec. 3) as an intermediate variable; in terms of z

$$x=\frac{2}{3}z^{\frac{3}{2}}. \tag{7.8}$$

In our first application of theorem B to eq (7.7) we make a cut in the z-plane along the part of the level curve of the function $\exp(\frac{2}{3}z^{\frac{3}{2}})$ which joins the point E of affix $-(\frac{3}{4}\pi)^{\frac{2}{3}}e^{\frac{4}{3}i\theta}$ and corresponding to $t=-1$, to the point at infinity on the negative real z-axis. Such a join can be made when $|\theta| \leq \frac{1}{2}\pi$ (cf. [7], fig. 2).

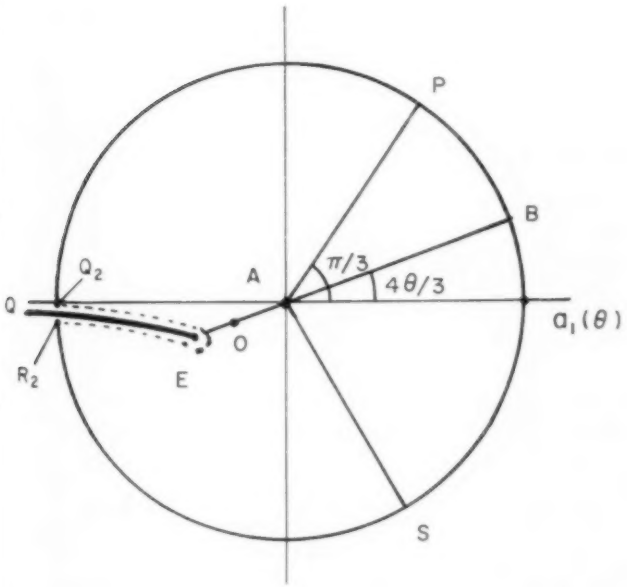


FIGURE 6. *z-plane: domains* $\mathbf{D}(\theta)$ *and* $\mathbf{G}(\theta)$ $\left(|\theta| \leq \frac{1}{2}\pi-\epsilon\right)$.

Figure 6 illustrates the mapping, the lettering corresponding to figures 2 and 3. The cut is the heavy curve EQ. The principal curves through A are mapped on the rays $\arg z=\pm\frac{1}{3}\pi$ and π; the principal curves through E are mapped on the two sides of the cut EQ and the remaining part of the level curve of $\exp(\frac{2}{3}z^{\frac{3}{2}})$ passing through E. Cut in this way, the z-plane is mapped on the complement of the domain $\mathbf{U}_2(\theta)$ (see fig. 4) in the t-plane.

When $|\theta| \leq \frac{1}{2}\pi-\epsilon$, we take the z-domain $\mathbf{D}(\theta)$ of [9], section 9, to be the cut plane with the circle center E and radius $\frac{1}{2}\delta$ removed, δ being the fixed small positive number introduced in

section 3. Then $F(e^{-\frac{2}{3}i\theta}z)$ is regular in $\mathbf{D}(\theta)$ (including $z=0$), and the condition (9.2) of [9] is satisfied with $\sigma=\frac{3}{2}$ because

$$e^{-\frac{5}{3}i\theta}F(e^{-\frac{2}{3}i\theta}z)\sim-\frac{7}{64z^2} \tag{7.9}$$

as $|z|\to\infty$, uniformly with respect to bounded arg z and bounded θ; this result follows from (7.3), (7.5), and (3.5).

We take the z-domain $\mathbf{G}(\theta)$ required by [9], section 9, to be the complement of the map of the t-domain $\mathbf{V}_2(\theta)$, defined in section 5. The boundaries of $\mathbf{G}(\theta)$ are the dotted curves surrounding the cut EQ in figure 6. All the preliminary conditions of [9], section 9, are then satisfied if $\delta=\delta(\epsilon)$ is taken to be small enough.

Taking the point $a_1(\theta)$ of [9], section 9 to be at infinity on the positive real axis, we find $\mathbf{H}_1(\theta)\equiv\mathbf{G}(\theta)$. Hence from theorem B, a solution W_1 of (7.7) exists, such that for large positive values of u

$$W_1\sim\mathrm{Ai}(u^{\frac{2}{3}}z)\sum_{s=0}^{\infty}\frac{A_s(\theta,z)}{u^{2s}}+\frac{\mathrm{Ai}'(u^{\frac{2}{3}}z)}{u^{\frac{1}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\theta,z)}{u^{2s}}\qquad\left(z\,\epsilon\,\mathbf{G}(\theta),\ \ |\theta|\leq\frac{1}{2}\pi-\epsilon\right)$$

uniformly with respect to z and θ, the coefficients $A_s(\theta,z)$ and $B_s(\theta,z)$ being given by recurrence relations derived from eqs (9.12) and (9.13) of [9]. When the variables μ and ζ are restored, this result becomes

$$W_1\sim\mathrm{Ai}(\mu^{\frac{1}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{3s}}+\frac{\mathrm{Ai}'(\mu^{\frac{1}{3}}\zeta)}{\mu^{\frac{5}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{3s}} \tag{7.10}$$

as $|\mu|\to\infty$, when

$$t\,\epsilon\,C\mathbf{V}_2(\theta)\text{ and }-\frac{1}{2}\pi+\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon, \tag{7.11}$$

where $A_0(\zeta)=$constant and

$$\left.\begin{aligned}B_s(\zeta)&=\frac{1}{2}\zeta^{-\frac{1}{2}}\int_0^{\zeta}\zeta^{-\frac{1}{2}}\{-A_s''(\zeta)+F(\zeta)A_s(\zeta)\}d\zeta\\A_{s+1}(\zeta)&=-\frac{1}{2}B_s'(\zeta)+\frac{1}{2}\int^{\zeta}F(\zeta)B_s(\zeta)d\zeta\end{aligned}\right\}\quad(s=0,1,2,\ldots), \tag{7.12}$$

the paths of integration lying in the map of $C\mathbf{U}_2(\theta)$.

A second application of theorem B can be made to the differential eq (7.7) by taking the cut in the z-plane for the function $F(e^{-\frac{2}{3}i\theta}z)$ to be the part of the level curve of $\exp(\frac{2}{3}z^{\frac{3}{2}})$ joining E to S, the point at infinity on arg $z=-\frac{1}{3}\pi$. This join is possible if $0\leq\theta\leq\pi$; the corresponding t-domain is then $C\mathbf{U}_3(\theta)$.

With similar analysis, again taking $a_1(\theta)$ to be the point $z=+\infty$, we may show that a solution of eq (7.2) exists having the right-hand side of (7.10) as its asymptotic expansion for large $|\mu|$, uniformly valid with respect to t and θ when

$$t\,\epsilon\,C\mathbf{V}_3(\theta)\text{ and }\epsilon\leq\theta\leq\pi-\epsilon. \tag{7.13}$$

The coefficients $A_s(\zeta)$ and $B_s(\zeta)$ are again given by (7.12), the paths of integration now lying in the map of $C\mathbf{U}_3(\theta)$.

By considering the limiting behavior as $\zeta\to\infty e^{-\frac{2}{3}i\theta}$, we see that when $\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon$, the ratio of the new solution to W_1 is independent of ζ and has an asymptotic expansion for large $|\mu|$ identically equal to unity. Hence the expansion (7.10) for W_1 holds with the conditions (7.13). Similar analysis, or an appeal to symmetry, shows that this expansion also holds with the further conditions

$$t\,\epsilon\,C\mathbf{V}_1(\theta)\text{ and }-\pi+\epsilon\leq\theta\leq-\epsilon. \tag{7.14}$$

We can combine the three t-regions of validity into a single region $\mathbf{T}(\theta)$, given by

$$\left.\begin{array}{lllll} C\mathbf{V}_1, & C\mathbf{V}_1 \cup C\mathbf{V}_2, & C\mathbf{V}_2, & C\mathbf{V}_2 \cup C\mathbf{V}_3, & C\mathbf{V}_3, \\ \text{according as } \theta \text{ lies in the respective intervals} & & & & \\ (-\pi+\epsilon, -\frac{1}{2}\pi+\epsilon), & (-\frac{1}{2}\pi+\epsilon, -\epsilon), & (-\epsilon, \epsilon), & (\epsilon, \frac{1}{2}\pi-\epsilon), & (\frac{1}{2}\pi-\epsilon, \pi-\epsilon). \end{array}\right\} \quad (7.15)$$

$\mathbf{T}(\theta)$ is illustrated in figure 7 for $0 \leq \theta \leq \pi-\epsilon$; $\mathbf{T}(-\theta)$ is the conjugate of $\mathbf{T}(\theta)$. The dotted curves are of the same character as the dotted curves of figure 2; their x-maps parallel the x-maps of the principal curves through E at a distance δ away. $\mathbf{T}(\theta)$ is the unshaded domain bounded by the dotted curves.

Collecting together the results of this section, we have shown that there exists a solution W_1 of the differential eq (7.2) whose asymptotic expansion for large $|\mu|$ is given by (7.10) and is uniformly valid with respect to $t \in \mathbf{T}(\theta)$ and θ in the interval $(-\pi+\epsilon, \pi-\epsilon)$. The functions $\zeta(t), F(\zeta)$, $A_s(\zeta)$ and $B_s(\zeta)$ are given by (7.4), (7.3), and (7.12), and are regular when $t \in \mathbf{T}(\theta)$.

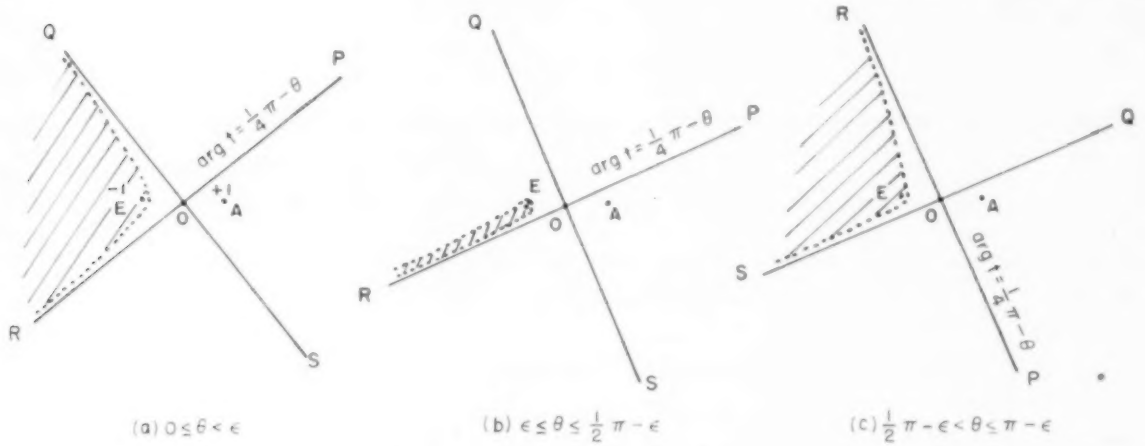


FIGURE 7. *t-plane: domains* $\mathbf{T}(\theta)$.

8. Identification of Solutions

From (4.1) and the second of eqs (7.1), we see that

$$W = \left(\frac{t^2-1}{\zeta}\right) U\left(-\frac{1}{2}\mu^2, \mu t\sqrt{2}\right), \quad (8.1)$$

satisfies the differential eq (7.2). When μ is fixed and $t \to \infty e^{-i\arg\mu}$ this solution becomes exponentially small. If $|\mu|$ is sufficiently large and $|\arg\mu| \leq \pi-\epsilon$, the solution W_1 of section 7 is also exponentially small in the same circumstances, for from (7.4) we have $\zeta \sim (\frac{3}{4})^{\frac{2}{3}} t^{\frac{4}{3}}$, and so $\mu^{\frac{4}{3}}\zeta \to +\infty$. Hence the two solutions are linearly dependent. Accordingly,

$$U\left(-\frac{1}{2}\mu^2, \mu t\sqrt{2}\right) \sim h(\mu)\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}\left[\mathrm{Ai}\,(\mu^{\frac{4}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}} + \frac{\mathrm{Ai}\,'(\mu^{\frac{4}{3}}\zeta)}{\mu^{\frac{8}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right] \quad (8.2)$$

as $|\mu| \to \infty$, uniformly with respect to $t \in \mathbf{T}(\arg\mu)$ and $\arg\mu$ in the interval $(-\pi+\epsilon, \pi-\epsilon)$.

This is the fundamental expansion in terms of Airy functions. In order to determine the function $h(\mu)$, we first fix the arbitrary constants of integration in the recurrence relations (7.12) by specifying the conditions

$$A_0(\zeta) = 1, \qquad \lim_{|\zeta|\to\infty} A_s(\zeta) = 0 \quad (s = 1, 2, \ldots). \quad (8.3)$$

We now examine the limiting form of the ratio of the two sides of the relation (8.2) as $t \to \infty e^{-i \arg \mu}$. With the aid of (2.2), (3.5), (7.5), (8.3), and the asymptotic forms

$$\mathrm{Ai}\,(\mu^{\frac{2}{3}}\zeta) \sim \tfrac{1}{2}\,\pi^{-\frac{1}{2}}(\mu^{\frac{2}{3}}\zeta)^{-\frac{1}{4}} e^{-\mu^2 \xi}, \qquad \mathrm{Ai}'(\mu^{\frac{2}{3}}\zeta) \sim -\tfrac{1}{2}\,\pi^{-\frac{1}{2}}(\mu^{\frac{2}{3}}\zeta)^{\frac{1}{4}} e^{-\mu^2 \xi}$$

(cf. [8], appendix), we find that $1/h(\mu)$ has the asymptotic expansion

$$\frac{1}{h(\mu)} \sim \pi^{-\frac{1}{2}} 2^{\frac{1}{4}\mu^2 - \frac{3}{4}}\, \mu^{-\frac{1}{2}\mu^2 + \frac{1}{4}}\, e^{\frac{1}{4}\mu^2} \left(1 - \sum_{s=0}^{\infty} \frac{h_s}{\mu^{4s+2}}\right) \qquad (|\arg \mu| \leq \pi - \epsilon) \tag{8.4}$$

as $|\mu| \to \infty$, uniformly valid with respect to $\arg \mu$, where

$$h_s = \lim_{|\zeta| \to \infty} \{\zeta^{\frac{1}{2}} B_s(\zeta)\}. \tag{8.5}$$

An alternative form of expansion, not involving the quantities h_s, may be found as follows. Suppose that $\epsilon \leq \theta \leq \frac{1}{2}\pi - \epsilon$. The principal curve AQ is then included in the domain $\mathbf{T}(\theta)$. On this curve $\mu^{\frac{2}{3}}\zeta$ is negative (cf. fig. 6). Let us set

$$\zeta = \zeta_m \equiv \mu^{-\frac{4}{3}} a'_m, \tag{8.6}$$

where a'_m is the mth negative zero of the function Ai′. Then (8.2) becomes

$$U\left(-\frac{1}{2}\,\mu^2, \mu t_m \sqrt{2}\right) \sim h(\mu) \left(\frac{\zeta_m}{t_m^2 - 1}\right)^{\frac{1}{4}} \mathrm{Ai}\,(a'_m) \sum_{s=0}^{\infty} \frac{A_s(\zeta_m)}{\mu^{2s}}, \tag{8.7}$$

in which t_m denotes the value of t corresponding to $\zeta = \zeta_m$. Now keep μ fixed and let $m \to \infty$, so that $t \to \infty e^{i(\frac{3}{4}\pi - \theta)}$. Then from [8], appendix, we have

$$a'_m = -\left\{\frac{3}{8}\,\pi(4m-3)\right\}^{\frac{2}{3}} \{1 + O(m^{-2})\}, \qquad \mathrm{Ai}\,(a'_m) = (-)^{m-1} \pi^{-\frac{1}{2}} (-a'_m)^{-\frac{1}{4}} \{1 + O(m^{-2})\}, \tag{8.8}$$

and from (7.5) and (3.5)

$$t_m^2 - \ln 2t_m = \frac{4}{3}\,\zeta_m^{\frac{3}{2}} + \frac{1}{2} + O(|t_m|^{-2}). \tag{8.9}$$

Substituting in (2.7) by means of these equations, we find that the limiting form of the left-hand side of (8.7) is minus the expression (5.17), with t_m defined of course as in the present section. On the other hand, the limiting form of the right of (8.7) is

$$h(\mu) \left(\frac{\mu^{-\frac{4}{3}} a'_m}{t_m^2 - 1}\right)^{\frac{1}{4}} (-)^{m-1} \pi^{-\frac{1}{2}} (-a'_m)^{-\frac{1}{4}} \sim (-)^{m-1} \pi^{-\frac{1}{2}} e^{i\frac{\pi}{6}} \mu^{-\frac{1}{3}}\, h(\mu)\, t_m^{-\frac{1}{2}}$$

(cf. (8.3) and (8.8)). Comparing the two sides we find that $\frac{1}{2}\,\pi^{-\frac{1}{2}}\,\mu^{-\frac{1}{3}}\,h(\mu)$ has the same asymptotic expansion for large $|\mu|$ as the right-hand side of the relation (5.18). Hence we obtain

$$h(\mu) \simeq 2\pi^{\frac{1}{2}}\,\mu^{\frac{1}{3}}\,\tilde{g}(\mu) \simeq 2\pi^{\frac{1}{2}}\,\mu^{\frac{1}{3}}\,g(\mu), \tag{8.10}$$

where $\tilde{g}(\mu)$, $g(\mu)$ are defined in sections 4 and 5; $g(\mu)$ has the asymptotic expansions (6.1) and (6.2) for large $|\mu|$. Although (8.10) has been proved on the assumption that $\epsilon \leq \arg \mu \leq \frac{1}{2}\pi - \epsilon$, (8.4) shows that $1/h(\mu)$, and hence also $h(\mu)$, possesses a single asymptotic expansion over the range $|\arg \mu| \leq \pi - \epsilon$. Therefore (8.10) must also hold over this wider range of $\arg \mu$.

Substituting (8.10) in (8.2), we obtain the final form of the fundamental expansion, given by

$$U\left(-\frac{1}{2}\,\mu^2, \mu t \sqrt{2}\right) \sim 2\pi^{\frac{1}{2}}\,\mu^{\frac{1}{3}}\,g(\mu) \left(\frac{\zeta}{t^2 - 1}\right)^{\frac{1}{4}} \left[\mathrm{Ai}\,(\mu^{\frac{2}{3}}\zeta) \sum_{s=0}^{\infty} \frac{A_s(\zeta)}{\mu^{4s}} + \frac{\mathrm{Ai}'(\mu^{\frac{2}{3}}\zeta)}{\mu^{\frac{2}{3}}} \sum_{s=0}^{\infty} \frac{B_s(\zeta)}{\mu^{4s}}\right] \tag{8.11}$$

as $|\mu| \to \infty$, uniformly with respect to $t \in \mathbf{T}(\arg \mu)$, $\arg \mu$ in the interval $(-\pi + \epsilon,\ \pi - \epsilon)$.

The coefficients $A_s(\zeta)$ and $B_s(\zeta)$ appearing in (8.11) are determined by the recurrence relations (7.12) and the conditions (8.3). Explicit expressions can be found by the method used in [8], section 6 for the determination of the coefficients in the uniform expansions of Bessel functions of large order. We take ζ to be fixed and positive, and let $\mu\to+\infty$. The Airy functions appearing in (8.11) can then be expanded in their asymptotic series for large positive arguments and the result compared with (4.3). Thus we derive the expressions

$$A_s(\zeta)=\sum_{m=0}^{2s} b_m\zeta^{-\frac{3}{2}m}\mathscr{A}_{2s-m}(\xi), \qquad \zeta^{\frac{1}{2}}B_s(\zeta)=-\sum_{m=0}^{2s+1} a_m\zeta^{-\frac{3}{2}m}\mathscr{A}_{2s-m+1}(\xi) \tag{8.12}$$

(cf. [8], (6.6)), where $\mathscr{A}_m(\xi)$ is defined in section 4, $a_0=1$ and

$$a_m=\frac{(2m+1)\ (2m+3)\ \ldots\ (6m-1)}{m!(144)^m}, \qquad b_m=-\frac{6m+1}{6m-1}\,a_m. \tag{8.13}$$

Analytic continuation shows that eqs (8.12) hold for all values of ζ.

We note from (8.10), (8.4), and (6.1) that

$$h_s\equiv\lim_{|\zeta|\to\infty}\{\zeta^{\frac{1}{2}}B_s(\zeta)\}=-g_{2s+1} \qquad (s=0,1,\ldots), \tag{8.14}$$

where g_s is given by (4.12) and (5.21).

Expansion for the derivative. The expansion (8.11) may be differentiated term by term with respect to t; this is a consequence of theorem B. With the aid of the first of eqs (7.1), we find that

$$U'\left(-\frac{1}{2}\mu^2,\,\mu t\sqrt{2}\right)\sim\frac{(2\pi)^{\frac{1}{2}}\mu^{\frac{2}{3}}g(\mu)}{\phi(\zeta)}\left[\frac{\mathrm{Ai}(\mu^{\frac{4}{3}}\zeta)}{\mu^{\frac{1}{3}}}\sum_{s=0}^{\infty}\frac{C_s(\zeta)}{\mu^{4s}}+\mathrm{Ai}'(\mu^{\frac{4}{3}}\zeta)\sum_{s=0}^{\infty}\frac{D_s(\zeta)}{\mu^{4s}}\right] \tag{8.15}$$

as $|\mu|\to\infty$, uniformly with respect to $t\in\mathbf{T}(\arg\mu)$ and $\arg\mu$ in the interval $(-\pi+\epsilon,\pi-\epsilon)$.

Here we have used the notation

$$\phi(\zeta)\equiv\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}=\left(\frac{dt}{d\zeta}\right)^{\frac{1}{2}}. \tag{8.16}$$

The coefficients $C_s(\zeta)$ and $D_s(\zeta)$ are expressed in terms of $A_s(\zeta)$ and $B_s(\zeta)$ by the equations

$$C_s(\zeta)=\chi(\zeta)A_s(\zeta)+A_s'(\zeta)+\zeta B_s(\zeta), \qquad D_s(\zeta)=A_s(\zeta)+\chi(\zeta)B_{s-1}(\zeta)+B_{s-1}'(\zeta), \tag{8.17}$$

(cf. [8], (6.10)), where

$$\chi(\zeta)\equiv\frac{\phi'(\zeta)}{\phi(\zeta)}=\frac{1}{4\zeta}-\frac{t}{2}\,\frac{\zeta^{\frac{1}{2}}}{(t^2-1)^{\frac{1}{2}}}=\frac{1-2t\{\phi(\zeta)\}^6}{4\zeta}. \tag{8.18}$$

Explicitly,

$$\zeta^{-\frac{1}{2}}C_s(\zeta)=-\sum_{m=0}^{2s+1} b_m\zeta^{-\frac{3}{2}m}\mathscr{B}_{2s-m+1}(\xi), \qquad D_s(\zeta)=\sum_{m=0}^{2s} a_m\zeta^{-\frac{3}{2}m}\mathscr{B}_{2s-m}(\xi), \tag{8.19}$$

(cf. [8], (6.12)), where $\mathscr{B}_m(\xi)$ is given by (4.14) to (4.16), and a_m, b_m by (8.13).

9. Use of Connection Formulas

In this section we seek expansions which hold in t-regions complementary to $\mathbf{T}(\arg\mu)$ when $|\arg\mu|\le\frac{1}{2}\pi$ (cf. sec 5).

In (8.11) replace μ by $\mu e^{i\pi}$. Then using (5.24) we obtain

$$U\left(-\frac{1}{2}\mu^2,-\mu t\sqrt{2}\right)\sim 2\pi^{\frac{1}{2}}e^{i\pi(\frac{1}{2}\mu^2-\frac{1}{6})}\mu^{\frac{1}{3}}g(\mu)\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}$$

$$\times\left[\mathrm{Ai}\,(e^{-\frac{2}{3}\pi i}\mu^{\frac{4}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(e^{-\frac{2}{3}\pi i}\mu^{\frac{4}{3}}\zeta)}{e^{\frac{2}{3}\pi i}\mu^{\frac{8}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right], \tag{9.1}$$

when

$$-\pi+\epsilon\leq\theta\leq-\epsilon, \qquad t\in\mathbf{T}(\theta+\pi). \tag{9.2}$$

In particular, we see from (7.15) and (5.3) that (9.1) holds when

$$-\frac{1}{2}\pi\leq\theta\leq-\epsilon, \qquad t\in C\mathbf{V}_3(\theta+\pi)=C\mathbf{V}_1(\theta). \tag{9.3}$$

Similarly,

$$U\left(-\frac{1}{2}\mu^2,-\mu t\sqrt{2}\right)\sim 2\pi^{\frac{1}{2}}e^{-i\pi(\frac{1}{2}\mu^2-\frac{1}{6})}\mu^{\frac{1}{3}}g(\mu)\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}$$
$$\times\left[\mathrm{Ai}\,(e^{\frac{2}{3}\pi i}\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(e^{\frac{2}{3}\pi i}\mu^{\frac{2}{3}}\zeta)}{e^{-\frac{2}{3}\pi i}\mu^{\frac{2}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right], \tag{9.4}$$

when

$$\epsilon\leq\theta\leq\frac{1}{2}\pi, \qquad t\in C\mathbf{V}_3(\theta). \tag{9.5}$$

The expansions (9.1) and (9.4) are of course the expansions for $U(-\frac{1}{2}\mu^2,\,\mu t\sqrt{2})$ in terms of Airy functions valid at the turning point $t=-1$. Combined with (8.11) they cover the whole of the t-plane when $\epsilon\leq|\theta|\leq\frac{1}{2}\pi$.

When $|\theta|\leq\epsilon$ we cannot expect to be able to represent the function $U(-\frac{1}{2}\mu^2,\,\mu t\sqrt{2})$ in $\mathbf{V}_2(\theta)=C\mathbf{T}(\theta)$ by means of a single Airy function, because as $|t|\to\infty$ in this region $U\,(-\frac{1}{2}\mu^2,\mu t\sqrt{2})$ becomes exponentially small or exponentially large according as μ^2 is or is not an odd positive integer (cf. (2.7)).

To deal with this remaining case we use the connection formula

$$U\left(-\frac{1}{2}\mu^2,\,-\mu t\sqrt{2}\right)-\sin\frac{1}{2}\pi\mu^2\cdot U\left(-\frac{1}{2}\mu^2,\,\mu t\sqrt{2}\right)=\left(\frac{\pi}{2}\right)^{\frac{1}{2}}\frac{e^{-i\pi(\frac{1}{4}\mu^2+\frac{1}{4})}}{\Gamma\,(\frac{1}{2}-\frac{1}{2}\mu^2)}U\left(\frac{1}{2}\mu^2,-i\mu t\sqrt{2}\right)$$
$$+\left(\frac{\pi}{2}\right)^{\frac{1}{2}}\frac{e^{i\pi(\frac{1}{4}\mu^2+\frac{1}{4})}}{\Gamma\,(\frac{1}{2}-\frac{1}{2}\mu^2)}U\left(\frac{1}{2}\mu^2,\,i\mu t\sqrt{2}\right), \tag{9.6}$$

obtained in an obvious manner from (5.26). Replacing μ by $\mu e^{-\frac{1}{2}i\pi}$ in (8.11) and using the first of eqs (5.25), we find

$$\left(\frac{\pi}{2}\right)^{\frac{1}{2}}\frac{e^{-i\pi(\frac{1}{4}\mu^2+\frac{1}{4})}}{\Gamma\,(\frac{1}{2}-\frac{1}{2}\mu^2)}U\left(\frac{1}{2}\mu^2,\,-i\mu t\sqrt{2}\right)\sim 2\pi^{\frac{1}{2}}e^{-\frac{1}{6}\pi i}\mu^{\frac{1}{3}}\cos\frac{1}{2}\pi\mu^2\cdot g(\mu)\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}$$
$$\times\left[\mathrm{Ai}\,(e^{-\frac{2}{3}\pi i}\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(e^{-\frac{2}{3}\pi i}\mu^{\frac{2}{3}}\zeta)}{e^{\frac{2}{3}\pi i}\mu^{\frac{2}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right].$$

We substitute this result and its conjugate form on the right of (9.6), and on the left we substitute (8.11). Using the eq [3]

$$\mathrm{Bi}\,(\mu^{\frac{2}{3}}\zeta)=e^{-\frac{1}{6}\pi i}\mathrm{Ai}\,(e^{-\frac{2}{3}\pi i}\mu^{\frac{2}{3}}\zeta)+e^{\frac{1}{6}\pi i}\mathrm{Ai}\,(e^{\frac{2}{3}\pi i}\mu^{\frac{2}{3}}\zeta),$$

and its differentiated form, we obtain

$$U\left(-\frac{1}{2}\mu^2,-\mu t\sqrt{2}\right)\sim 2\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}g(\mu)\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}\left[\sin\frac{1}{2}\pi\mu^2\left\{\mathrm{Ai}\,(\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(\mu^{\frac{2}{3}}\zeta)}{\mu^{\frac{2}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right\}\right.$$
$$\left.+\cos\frac{1}{2}\pi\mu^2\left\{\mathrm{Bi}\,(\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Bi}'(\mu^{\frac{2}{3}}\zeta)}{\mu^{\frac{2}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right\}\right]. \tag{9.7}$$

This is valid when $|\theta|\leq\frac{1}{2}\pi-\epsilon$ in the t-region common to the domains $\mathbf{T}(\theta)$, $\mathbf{T}(\theta-\frac{1}{2}\pi)$, and

$\mathbf{T}(\theta+\frac{1}{2}\pi)$. Using (7.15) and (5.2), we see that this means (9.7) is valid when

$$|\theta|\leq\frac{1}{2}\pi-\epsilon, \qquad t\,\epsilon\, C\mathbf{V}_2(\theta). \tag{9.8}$$

This is the required result. We note that when μ^2 is an odd integer the terms in Bi, Bi′ in (9.7) vanish, as we require from a consideration of the behavior of $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ as $|t|\to\infty$ in the sector $\mathbf{V}_2(\theta)$. We may also verify without difficulty that (9.7) agrees with (9.1) and (9.4) in their common t-regions of validity when $-\frac{1}{2}\pi+\epsilon\leq\theta\leq-\epsilon$ and $\epsilon\leq\theta\leq\frac{1}{2}\pi-\epsilon$, respectively.

10. Summary of Expansions in Terms of Airy Functions

The fundamental expansion of this kind is (8.11). It is valid for large $|\mu|$ uniformly with respect to $\theta\equiv\arg\mu$ and t, when $-\pi+\epsilon\leq\theta\leq\pi-\epsilon$ and t lies in the domain $\mathbf{T}(\theta)$, defined by (7.15) and illustrated in figure 7. The function $g(\mu)$ is calculable from either of the asymptotic expansions (6.1) and (6.2). The function $\zeta(t)$ is given by (7.4), that branch being chosen which is continuous in $\mathbf{T}(\theta)$ and positive when $t>1$; similarly the branch of $\{\zeta/(t^2-1)\}^{\frac{1}{4}}$ is continuous in $\mathbf{T}(\theta)$ and positive when $t>1$. The coefficients $A_s(\zeta)$ and $B_s(\zeta)$ are given by eqs (8.12); they satisfy the recurrence relations (7.12) and the eqs (8.3). They are regular functions of ζ in the ζ-map of $\mathbf{T}(\theta)$ and real when ζ lies on the real axis to the right of the point of affix $-(\frac{3}{4}\pi)^{\frac{2}{3}}$.

Other asymptotic expansions for $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ are (9.1), (9.4), and (9.7); they are valid with the conditions (9.3), (9.5), and (9.8), respectively. The combination of these expansions and (8.11) covers the whole t-plane when $|\theta|\leq\frac{1}{2}\pi$.

The expansion for the derivative $U'(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ corresponding to (8.11) is (8.15), and holds under the same conditions. The coefficients $C_s(\zeta)$ and $D_s(\zeta)$ are given by (8.17) and (8.19). The derivative expansions corresponding to (9.1), (9.4), and (9.7) are not recorded, but they can be obtained immediately by analogy with (8.15). They are valid under the same conditions as the parent formulas.

Approximations for parabolic cylinder functions of large orders which hold in regions containing one of the turning points have been given by Watson ([12], sec. 17), Schwid ([10], sec. 6), Erdélyi, Kennedy, and McGregor [2], and Kazarinoff [15].

Watson derives asymptotic expansions in terms of elementary functions for the functions $D_n(z\pm2\sqrt{n})\equiv U(-n-\frac{1}{2}, z\pm2\sqrt{n})$ as $|n|\to\infty$, with the conditions $|\arg n|<\frac{1}{2}\pi$ and $z=o(|n|^{-\frac{1}{6}})$. In the present notation these conditions correspond to $|\arg\mu|<\frac{1}{4}\pi$ and $t\mp1=o(|\mu|^{-\frac{2}{3}})$. Watson's expansions are analogous to Meissel's expansions for Bessel functions of large order and may be obtained from the present uniform expansions by suitable re-expansion (cf. [8], sec. 6).

Schwid gives approximations in terms of Bessel functions of order one-third (Airy functions) which are valid in the neighborhood $t-1=O(|\mu|^{-2})$ as $|\mu|\to\infty$, using the present notation. This neighborhood is complementary to the region of validity of Schwid's approximations in terms of elementary functions, described in section 6. No investigation is made of the validity of the Bessel function approximations outside the neighborhood.

The approach of Erdélyi, Kennedy, and McGregor is similar to the one we have used. These writers obtain a set of asymptotic approximations in terms of Airy functions, to solutions of eq (1.1) when μ and t are both complex (using the present notation). The results are contained in those given in the present paper; the principal ways in which they have been extended and improved upon are as follows.

First, the approximations apply to fixed values of $\arg\mu$, no investigation is made of their uniform validity with respect to $\arg\mu$.

Secondly, the approximations are established with the condition $|\arg\mu|\leq\frac{1}{4}\pi$; outside this interval connection formulas are to be used. This leads to complexity in the number of results; eleven are given compared with the four we have found to be sufficient.

Finally, only the leading terms of the asymptotic expansions are found; no terms involving the first derivative of the Airy functions appear in the approximations. An effect of this is

to exclude regions in the t-plane containing the zeros of the Airy functions. No special treatment of such regions is required in the present theory.

Kazarinoff's paper (which has appeared since the present paper was written) contains a brief investigation of the extension of the results of Erdélyi, Kennedy, and McGregor to include uniformity with respect to arg μ and higher terms in the asymptotic expansions. The results are rather more complicated and less complete than those given here.

Part 4. Results for Real Variables

In this part we suppose μ, t, a, and x to denote real variables except where otherwise indicated; μ is always large and positive. The variable x is not to be confused with the complex function $x(t)$ introduced in parts 2 and 3.

11. The Equation $w''=(\frac{1}{4}x^2+a)w$

The principal solutions of this equation are $U(a,x)$, $U(a,-x)$, $V(a,x)$, and $\overline{U}(a,x)$, defined in section 2.

(i) *a positive.* When $a \to +\infty$ the differential equation has no turning points on the real x-axis. The expansions of greatest interest are therefore those in terms of elementary functions, and we shall not consider the forms of the Airy expansions.

In (4.3) replace μ,t by $i\mu$, $-it$, respectively. Then using (4.8), we obtain

$$U\left(\frac{1}{2}\mu^2, \mu t\sqrt{2}\right) \sim g(\mu e^{\frac{1}{2}i\pi}) \frac{e^{\mu^2\xi(-it)}}{(-t^2-1)^{\frac{1}{4}}} \sum_{s=0}^{\infty} (-)^s \frac{u_s(-it)}{(-t^2-1)^{\frac{1}{2}s}} \frac{1}{\mu^{2s}} \tag{11.1}$$

as $\mu \to +\infty$. In the complex plane the region of validity of this expansion is $e^{\frac{1}{2}i\pi}\mathbf{S}(\frac{1}{2}\pi)$, and from figure 2(c) it is clear that this region includes the whole of the real t-axis.

We have now to interpret the branches of the various functions.

From (6.1) and (6.2) we derive

$$g(\mu e^{\frac{1}{2}i\pi}) = e^{-i\pi(\frac{1}{4}+\frac{1}{4}\mu^2)}\overline{g}(\mu), \tag{11.2}$$

where the function $\overline{g}(\mu)$ is real and has the expansions

$$\frac{1}{\overline{g}(\mu)} \sim 2^{-\frac{1}{4}\mu^2+\frac{1}{4}} e^{-\frac{1}{4}\mu^2} \mu^{\frac{1}{2}\mu^2+\frac{1}{2}} \left\{1 - \sum_{s=0}^{\infty} \frac{g_{2s+1}}{\mu^{4s+2}}\right\}, \tag{11.3}$$

$$\overline{g}(\mu) \sim 2^{\frac{1}{4}\mu^2-\frac{1}{4}} e^{\frac{1}{4}\mu^2} \mu^{-\frac{1}{2}\mu^2-\frac{1}{2}} \left\{1 + \frac{1}{2}\sum_{s=1}^{\infty} (-)^s \frac{\gamma_s}{(\frac{1}{2}\mu^2)^s}\right\}, \tag{11.4}$$

as $\mu \to +\infty$. Comparison of (11.4) with (2.27) shows incidentally that

$$\overline{g}(\mu) \simeq 2^{\frac{1}{4}\mu^2-\frac{3}{4}} e^{\frac{1}{4}\mu^2} \mu^{-\frac{1}{2}\mu^2-\frac{1}{2}} + \frac{\pi^{\frac{1}{2}} 2^{-\frac{1}{4}\mu^2-\frac{3}{4}} e^{-\frac{1}{4}\mu^2} \mu^{\frac{1}{2}\mu^2-\frac{1}{2}}}{\Gamma(\frac{1}{2}+\frac{1}{2}\mu^2)}, \tag{11.5}$$

where the symbol $\simeq$ is defined in section 5.

Next, when t lies on the negative imaginary axis of figure 2(c), $\xi(t)$ lies on the straight line $O'C'$ shown in the third diagram of figure 1. Hence

$$\xi(-it) = \frac{1}{4}i\pi - \overline{\xi}(t), \tag{11.6}$$

where $\overline{\xi}(t)$ is positive when $t>0$. From (3.4) we find

$$\overline{\xi}(t) = \int_0^t (t^2+1)^{\frac{1}{2}} dt = \frac{1}{2}t(t^2+1)^{\frac{1}{2}} + \frac{1}{2}\ln\{t+(t^2+1)^{\frac{1}{2}}\}, \tag{11.7}$$

the branches here having their principal values.

Lastly, we have

$$(-t^2-1)^{\frac{1}{4}}=e^{-\frac{1}{4}i\pi}(t^2+1)^{\frac{1}{4}}, \qquad (-t^2-1)^{\frac{1}{2}s}=e^{-\frac{1}{2}si\pi}(t^2+1)^{\frac{1}{2}s}, \tag{11.8}$$

where the functions on the right have their principal values.

Substituting in (11.1) by means of (11.2), (11.6), (11.8), and the relation

$$\bar{u}_s(t)\equiv i^s u_s(-it), \tag{11.9}$$

we obtain the required result in real form, given by

$$U\left(\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim\bar{g}(\mu)\frac{e^{-\mu^2\bar{\xi}(t)}}{(t^2+1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}\frac{\bar{u}_s(t)}{(t^2+1)^{\frac{1}{2}s}}\frac{1}{\mu^{2s}} \qquad (-\infty<t<\infty), \tag{11.10}$$

uniformly with respect to t. Explicit expressions for the coefficients $\bar{u}_s(t)$, obtained from (4.11), are

$$\left.\begin{aligned}&\bar{u}_0(t)=1, \qquad \bar{u}_1(t)=(-t^3-6t)/24, \qquad \bar{u}_2(t)=(9t^4+249t^2-145)/1152,\\ &\bar{u}_3(t)=(4042t^9+18189t^7+28287t^5-1\ 51995t^3+2\ 59290t)/4\ 14720.\end{aligned}\right\} \tag{11.11}$$

The corresponding result for the derivative is

$$U'\left(\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim-\frac{\mu}{\sqrt{2}}\,\bar{g}(\mu)(t^2+1)^{\frac{1}{4}}e^{-\mu^2\bar{\xi}(t)}\sum_{s=0}^{\infty}\frac{\bar{v}_s(t)}{(t^2+1)^{\frac{1}{2}s}}\frac{1}{\mu^{2s}}, \tag{11.12}$$

valid with the same conditions. Here

$$\bar{v}_s(t)\equiv i^s v_s(-it). \tag{11.13}$$

Explicitly,

$$\left.\begin{aligned}&\bar{v}_0(t)=1, \qquad \bar{v}_1(t)=(-t^3+6t)/24, \qquad \bar{v}_2(t)=(-15t^4-327t^2+143)/1152,\\ &\bar{v}_3(t)=(4042t^9+18189t^7+36387t^5+2\ 38425t^3-2\ 59290t)/4\ 14720,\end{aligned}\right\} \tag{11.14}$$

(cf. (4.13) to (4.16)).

Since $U(a,x)$ and $U(a,-x)$ comprise a satisfactory pair of solutions of the differential equation when a is positive, it is unnecessary to give the corresponding expansions for the solutions $V(a,x)$ and $\overline{U}(a,x)$.

(ii) *a negative; expansions in terms of elementary functions.* For $t\geq 1+\delta$, the expansion of $U(-\frac{1}{2}\mu^2,\ \mu t\sqrt{2})$ is given in real form by (4.3)[3]. In this expansion $g(\mu)$ is calculable from (6.1), (6.2), or the relation

$$g(\mu)\simeq 2^{-\frac{1}{4}\mu^2-\frac{1}{4}}e^{-\frac{1}{4}\mu^2}\mu^{\frac{1}{2}\mu^2-\frac{1}{2}}+\pi^{-\frac{1}{2}}2^{\frac{1}{4}\mu^2-\frac{1}{4}}e^{\frac{1}{4}\mu^2}\mu^{-\frac{1}{2}\mu^2-\frac{1}{2}}\Gamma\left(\frac{1}{2}+\frac{1}{2}\mu^2\right), \tag{11.15}$$

and ξ, $\mathscr{A}_s(\xi)$, by use of (3.4), (4.8), and (4.11), all fractional powers now taking their principal values.

When $-1+\delta\leq t\leq 1-\delta$, the expansion of $U(-\frac{1}{2}\mu^2,\ \mu t\sqrt{2})$ is given by (5.11). Here $\bar{g}(\mu)$ is identical with $g(\mu)$ and η is given by (5.9).

For the remaining part of the real axis we have the expansion (5.28). In the present circumstances the function $\xi(t)$ is positive and bounded away from zero. Hence the contribution of the series $i\cos\frac{1}{2}\pi\mu^2\cdot e^{-\mu^2\xi}\sum\mathscr{A}_s(\xi)\mu^{-2s}$ is always asymptotically negligible compared with that of $2\cos\frac{1}{2}\pi\mu^2\cdot e^{\mu^2\xi}\sum(-)^s\mathscr{A}_s(\xi)\mu^{-2s}$. Thus (5.28) reduces to

$$U\left(-\tfrac{1}{2}\mu^2,-\mu t\sqrt{2}\right)\sim\frac{g(\mu)}{(t^2-1)^{\frac{1}{4}}}\left[2\cos\frac{1}{2}\pi\mu^2\cdot e^{\mu^2\xi}\sum_{s=0}^{\infty}(-)^s\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}+\sin\frac{1}{2}\pi\mu^2\cdot e^{-\mu^2\xi}\sum_{s=0}^{\infty}\frac{\mathscr{A}_s(\xi)}{\mu^{2s}}\right], \tag{11.16}$$

[3] For convenience the symbol δ' of parts 2 and 3, denoting an arbitrary fixed positive number, is replaced here by δ.

valid when $t \geq 1+\delta$. The functions $g(\mu)$, ξ and $\mathscr{A}_s(\xi)$ are the same as in (4.3), discussed above.

The corresponding expansions for the second solution $\overline{U}(a, x)$ are of interest, particularly when $x \geq 0$. It is not possible to obtain them trivially by substituting in the connection formula (2.12) the expansions we have just obtained for $U(a, x)$ and $U(a, -x)$, because although cancellation of terms in the asymptotic series takes place, we cannot assert immediately that the remainder terms implied in these series also cancel. To put the results on a sound basis we substitute the expansions (5.5) and (5.6) in the formula

$$\overline{U}\left(-\tfrac{1}{2}\mu^2, \mu t\sqrt{2}\right) = (2\pi)^{-\frac{1}{2}}\,\Gamma\left(\tfrac{1}{2}+\tfrac{1}{2}\mu^2\right)\left[e^{-i\pi(\frac{1}{4}+\frac{1}{4}\mu^2)}\,U\left(\tfrac{1}{2}\mu^2, -i\mu t\sqrt{2}\right)+e^{i\pi(\frac{1}{4}+\frac{1}{4}\mu^2)}\,U\left(\tfrac{1}{2}\mu^2, i\mu t\sqrt{2}\right)\right], \tag{11.17}$$

obtained from (2.5) and (2.12).

When $t \geq 1+\delta$ the branches of $\xi(t)$ and $\mathscr{A}_s(\xi)$ in both expansions are the principal ones. With the aid of the relation

$$(2\pi)^{-\frac{1}{2}}\,\Gamma\left(\tfrac{1}{2}+\tfrac{1}{2}\mu^2\right) e^{\mp i\pi(\frac{1}{4}+\frac{1}{4}\mu^2)}\, g(\mu e^{\mp\frac{1}{2}i\pi}) \simeq g(\mu), \tag{11.18}$$

proved by use of (2.22), (5.7), (5.20), and (6.2), we deduce that

$$\overline{U}\left(-\tfrac{1}{2}\mu^2, \mu t\sqrt{2}\right) \sim 2g(\mu)\,\frac{e^{\mu^2\xi}}{(t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}(-)^s\,\frac{u_s(t)}{(t^2-1)^{\frac{3}{2}s}}\,\frac{1}{\mu^{2s}} \qquad (t \geq 1+\delta). \tag{11.19}$$

When $-1+\delta \leq t \leq 1-\delta$, we interpret the branches of the many-valued functions by use of (5.8) and (5.10). With the aid of (11.18) we obtain on reduction

$$\overline{U}\left(-\tfrac{1}{2}\mu^2, \mu t\sqrt{2}\right) \sim \frac{2g(\mu)}{(1-t^2)^{\frac{1}{4}}}\left[\cos(\mu^2\eta+\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s}(t)}{(1-t^2)^{3s}\mu^{4s}} - \sin(\mu^2\eta+\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s+1}(t)}{(1-t^2)^{3s+\frac{3}{2}}\mu^{4s+2}}\right], \tag{11.20}$$

valid when $-1+\delta \leq t \leq 1-\delta$.

The expansions for the derivatives corresponding to the expansions obtained in this subsection can be written down immediately by use of (6.3), or by analogy with (5.23).

(iii) *a negative; expansions in terms of Airy functions.* For $t \geq -1+\delta$ the expansions of $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ and $U(-\frac{1}{2}\mu^2, -\mu t\sqrt{2})$ are given by (8.11) and (9.7), respectively. In these expansions $\zeta(t)$ is given by (7.4) and is real when $t > -1$; a more convenient form of this relation when $-1 < t < 1$ is

$$\frac{2}{3}(-\zeta)^{\frac{3}{2}} = \int_t^1 (1-t^2)^{\frac{1}{2}}dt = \frac{1}{2}\cos^{-1}t - \frac{1}{2}\,t(1-t^2)^{\frac{1}{2}} = \eta(t) \tag{11.21}$$

(cf. (5.9)). The coefficients $A_s(\zeta)$ and $B_s(\zeta)$ are real and given by (8.12), (8.13), and (4.8)[4].

An expansion for the solution $\overline{U}(a,x)$ may be derived by use of (11.17), as in the case of the exponential-type expansions. The result is

$$\overline{U}\left(-\tfrac{1}{2}\mu^2, \mu t\sqrt{2}\right) \sim 2\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}g(\mu)\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}\left[\mathrm{Bi}(\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{\mu^{4s}} + \frac{\mathrm{Bi}'(\mu^{\frac{2}{3}}\zeta)}{\mu^{\frac{2}{3}}}\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{\mu^{4s}}\right], \tag{11.22}$$

valid when $t \geq -1+\delta$.

The corresponding expansion for $\overline{U}'(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ may be written down by analogy with (8.15).

(iv) *Hermite polynomials of large degree.* Expansions for the Hermite polynomial $H_n(x)$

[4] More convenient expressions when $-1 < t < 1$ are given by (13.4) and (13.5).

when n is large can be obtained from the results in subsections (ii) and (iii) above by use of the equation

$$H_{\frac{1}{2}\mu^2-\frac{1}{2}}(\mu t)=2^{\frac{1}{4}\mu^2-\frac{1}{4}}e^{\frac{1}{8}\mu^2t^2}U\left(-\frac{1}{2}\,\mu^2,\mu t\sqrt{2}\right). \tag{11.23}$$

Thus (4.3) and (5.11) yield expansions which hold uniformly in the ranges $x\geq(1+\delta)\,(2n+1)^{\frac{1}{2}}$ and $-(1-\delta)(2n+1)^{\frac{1}{2}}\leq x\leq(1-\delta)(2n+1)^{\frac{1}{2}}$, respectively. An expansion uniformly valid throughout the range $x\geq-(1-\delta)(2n+1)^{\frac{1}{2}}$ is furnished by the Airy expansion (8.11).

12. The Equation $w''=(a-\frac{1}{4}x^2)w$

(i) *a positive; expansions in terms of elementary functions.* From (2.16), (2.17) and (2.18), we have

$$k^{-\frac{1}{2}}W\left(\frac{1}{2}\,\mu^2,\mu t\sqrt{2}\right)+ik^{\frac{1}{2}}W\left(\frac{1}{2}\,\mu^2,-\mu t\sqrt{2}\right)=2^{\frac{1}{2}}e^{\frac{1}{8}\pi\mu^2}e^{i(\frac{1}{2}\phi_2+\frac{1}{8}\pi)}U\left(\frac{1}{2}\,i\mu^2,e^{-\frac{1}{4}\pi i}\mu t\sqrt{2}\right), \tag{12.1}$$

where

$$k=(1+e^{\pi\mu^2})^{\frac{1}{2}}-e^{\frac{1}{2}\pi\mu^2},\qquad e^{\frac{1}{2}i\phi_2}=\pm\left\{\Gamma\left(\frac{1}{2}+\frac{1}{2}\,i\mu^2\right)\Big/\left|\Gamma\left(\frac{1}{2}+\frac{1}{2}\,i\mu^2\right)\right|\right\}^{\frac{1}{2}}. \tag{12.2}$$

For large μ,

$$k=\frac{1}{2}\,e^{-\frac{1}{2}\pi\mu^2}+O(e^{-\frac{3}{2}\pi\mu^2}),\qquad e^{\frac{1}{2}i\phi_2}=e^{-\frac{1}{4}i\mu^2}\left(\frac{1}{2}\,\mu^2\right)^{\frac{1}{4}i\mu^2}[1+O(\mu^{-2})]. \tag{12.3}$$

In (4.3) replace μ by $\mu e^{-\frac{1}{4}\pi i}$. Then we obtain

$$U\left(\frac{1}{2}\,i\mu^2,e^{-\frac{1}{4}\pi i}\mu t\sqrt{2}\right)\sim g(\mu e^{-\frac{1}{4}\pi i})\,\frac{e^{i\mu^2\xi}}{(t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}i^s\,\frac{u_s(t)}{(t^2-1)^{\frac{1}{2}s}}\,\frac{1}{\mu^{2s}} \tag{12.4}$$

as $\mu\to+\infty$. The t-region of validity of this expansion, $\mathbf{S}(-\frac{1}{4}\pi)$, includes the segments $t\geq1+\delta$ and $-1+\delta\leq t\leq1-\delta$ of the real axis (cf. fig. 2(b)).

When $\mu\to+\infty$, we have from (12.2) and (2.22)

$$e^{\frac{1}{2}i\phi_2}\sim e^{-\frac{1}{4}i\mu^2}\left(\frac{1}{2}\,\mu^2\right)^{\frac{1}{4}i\mu^2}\left[\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\,i\mu^2)^s}\right]^{\frac{1}{2}}\left|\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\,i\mu^2)^s}\right|^{-\frac{1}{2}}, \tag{12.5}$$

the choice of ambiguous sign on taking the square root being resolved by use of (12.3). From (6.1)

$$g(\mu e^{-\frac{1}{4}\pi i})\sim2^{\frac{1}{4}i\mu^2-\frac{1}{4}}e^{\frac{1}{4}i\mu^2}e^{\frac{1}{8}\pi i-\frac{1}{8}\pi\mu^2}\mu^{-\frac{1}{2}i\mu^2-\frac{1}{2}}(1+iL)^{-1}, \tag{12.6}$$

where the symbol L denotes the asymptotic series

$$L\equiv\sum_{s=0}^{\infty}(-)^s\,\frac{g_{2s+1}}{\mu^{4s+2}}. \tag{12.7}$$

Further from (5.21), with μ replaced by $\mu e^{\frac{1}{4}\pi i}$, we have

$$\sum_{s=0}^{\infty}\frac{\gamma_s}{(\frac{1}{2}\,i\mu^2)^s}\sim\frac{2}{1-iL}-1=\frac{1+iL}{1-iL}. \tag{12.8}$$

Introducing a function $l(\mu)$, defined by

$$l(\mu)\equiv2^{\frac{1}{2}}e^{\frac{1}{8}\pi\mu^2}e^{i(\frac{1}{2}\phi_2-\frac{1}{8}\pi)}g(\mu e^{-\frac{1}{4}\pi i}), \tag{12.9}$$

and using (12.5), (12.6) and (12.8), we find that

$$l(\mu)\sim\frac{2^{\frac{1}{4}}}{\mu^{\frac{1}{2}}}\left(\frac{1+iL}{1-iL}\right)^{\frac{1}{2}}\left|\frac{1+iL}{1-iL}\right|^{-\frac{1}{2}}\frac{1}{1+iL}\sim\frac{2^{\frac{1}{4}}}{\mu^{\frac{1}{2}}}\,\frac{1}{(1+L^2)^{\frac{1}{2}}}. \tag{12.10}$$

Thus $l(\mu)$ is a real function, having an asymptotic expansion of the form

$$l(\mu)\sim\frac{2^{\frac{1}{4}}}{\mu^{\frac{1}{2}}}\sum_{s=0}^{\infty}\frac{l_s}{\mu^{4s}}\quad\text{as }\mu\to+\infty. \tag{12.11}$$

From (12.10), (12.7), and (4.12), we find

$$l_0=1,\qquad l_1=-\frac{1}{2}\,g_1^2=-\frac{1}{1152},\qquad l_2=g_1g_3+\frac{3}{8}\,g_1^4=-\frac{16123}{398\;13120}. \tag{12.12}$$

We are now in a position to combine (12.1) with (12.4). When $t\geq 1+\delta$ each of the functions $\xi(t)$, $(t^2-1)^{\frac{1}{4}}$ and $(t^2-1)^{\frac{3}{2}s}$ is positive. Substituting by means of (12.9) and the first of (12.3), and equating real and imaginary parts, we find

$$W(\tfrac{1}{2}\mu^2,\mu t\sqrt{2})\sim\frac{2^{-\frac{1}{2}}e^{-\frac{1}{4}\pi\mu^2}l(\mu)}{(t^2-1)^{\frac{1}{4}}}\left[\cos(\mu^2\xi+\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s}(t)}{(t^2-1)^{3s}\mu^{4s}}-\sin(\mu^2\xi+\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s+1}(t)}{(t^2-1)^{3s+\frac{3}{2}}\mu^{4s+2}}\right], \tag{12.13}$$

$$W(\tfrac{1}{2}\mu^2,-\mu t\sqrt{2})\sim\frac{2^{\frac{1}{2}}e^{\frac{1}{4}\pi\mu^2}l(\mu)}{(t^2-1)^{\frac{1}{4}}}\left[\sin(\mu^2\xi+\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s}(t)}{(t^2-1)^{3s}\mu^{4s}}+\cos(\mu^2\xi+\tfrac{1}{4}\pi)\sum_{s=0}^{\infty}\frac{(-)^s u_{2s+1}(t)}{(t^2-1)^{3s+\frac{3}{2}}\mu^{4s+2}}\right], \tag{12.14}$$

uniformly with respect to t when $t\geq 1+\delta$. The function $l(\mu)$ is calculable from (12.11).

When $-1+\delta\leq t\leq 1-\delta$ the appropriate branches of ξ and $(t^2-1)^{\frac{1}{4}}$ are given by $\xi=-i\eta$ and $(t^2-1)^{\frac{1}{4}}=e^{\frac{1}{4}\pi i}(1-t^2)^{\frac{1}{4}}$, where η is the real function (5.9) and $(1-t^2)^{\frac{1}{4}}$ is positive. Hence we derive

$$W\left(\frac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim\frac{l(\mu)}{2^{\frac{1}{2}}e^{\frac{1}{4}\pi\mu^2}}\frac{e^{\mu^2\eta}}{(1-t^2)^{\frac{1}{4}}}\sum_{s=0}^{\infty}\frac{(-)^s u_s(t)}{(1-t^2)^{\frac{3}{2}s}}\frac{1}{\mu^{2s}}\qquad(-1+\delta\leq t\leq 1-\delta). \tag{12.15}$$

The expansions for the derivatives corresponding to (12.13), (12.14), and (12.15) may be obtained by term-by-term differentiation with respect to t (cf. (5.23) and (4.13)).

It may be noted that (12.13), (12.14), and (12.15) are respectively equivalent to the formal series (317), (318) and (324) given on pages 84 and 85 of [4] (cf. sec. 6).

(ii) *a positive; expansions in terms of Airy functions.* In (8.11) replace μ by $\mu e^{-\frac{1}{4}\pi i}$. Then we obtain

$$U\left(\frac{1}{2}\,i\mu^2,e^{-\frac{1}{4}\pi i}\mu t\sqrt{2}\right)\sim\frac{2\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}}{e^{\frac{1}{12}\pi i}}\,g(\mu e^{-\frac{1}{4}\pi i})\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}$$

$$\times\left[\mathrm{Ai}(e^{-\frac{1}{3}\pi i}\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}(-)^s\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(e^{-\frac{1}{3}\pi i}\mu^{\frac{2}{3}}\zeta)}{e^{-\frac{2}{3}\pi i}\mu^{\frac{5}{3}}}\sum_{s=0}^{\infty}(-)^s\frac{B_s(\zeta)}{\mu^{4s}}\right]. \tag{12.16}$$

The t-region of validity of this expansion, $\mathbf{T}(-\frac{1}{4}\pi)$, includes the segment $t\geq -1+\delta$ (cf. fig. 7(b)). Now from [3]

$$2e^{\frac{1}{6}\pi i}\,\mathrm{Ai}\,(e^{-\frac{1}{3}\pi i}\mu^{\frac{2}{3}}\zeta)=\mathrm{Bi}\,(-\mu^{\frac{2}{3}}\zeta)+i\,\mathrm{Ai}\,(-\mu^{\frac{2}{3}}\zeta). \tag{12.17}$$

Hence substituting (12.16) in (12.1) and equating real parts, we find, with the aid of (12.9) and the first of (12.3),

$$W\left(\frac{1}{2}\,\mu^2,\mu t\sqrt{2}\right)\sim\frac{\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}l(\mu)}{2^{\frac{1}{2}}e^{\frac{1}{4}\pi\mu^2}}\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}\left[\mathrm{Bi}(-\mu^{\frac{2}{3}}\zeta)\sum_{s=0}^{\infty}(-)^s\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Bi}'(-\mu^{\frac{2}{3}}\zeta)}{\mu^{\frac{5}{3}}}\sum_{s=0}^{\infty}(-)^s\frac{B_s(\zeta)}{\mu^{4s}}\right] \tag{12.18}$$

valid when $t\geq -1+\delta$. The corresponding expansion for the derivative is

$$W'\left(\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim-\frac{\pi^{\frac{1}{2}}\mu^{\frac{2}{3}}l(\mu)}{2e^{\frac{1}{4}\pi\mu^2}}\left(\frac{t^2-1}{\zeta}\right)^{\frac{1}{4}}$$

$$\times\left[-\frac{\mathrm{Bi}(-\mu^{\frac{4}{3}}\zeta)}{\mu^{\frac{2}{3}}}\sum_{s=0}^{\infty}(-)^s\frac{C_s(\zeta)}{\mu^{4s}}+\mathrm{Bi}'(-\mu^{\frac{4}{3}}\zeta)\sum_{s=0}^{\infty}(-)^s\frac{D_s(\zeta)}{\mu^{4s}}\right] \quad (12.19)$$

(cf. (8.15)).

The expansion for the second solution may be obtained by equating imaginary parts. This gives

$$W\left(\tfrac{1}{2}\mu^2,-\mu t\sqrt{2}\right)\sim\frac{\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}l(\mu)}{2^{-\frac{1}{2}}e^{-\frac{1}{4}\pi\mu^2}}\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}\left[\mathrm{Ai}(-\mu^{\frac{4}{3}}\zeta)\sum_{s=0}^{\infty}(-)^s\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(-\mu^{\frac{4}{3}}\zeta)}{\mu^{\frac{5}{3}}}\sum_{s=0}^{\infty}(-)^s\frac{B_s(\zeta)}{\mu^{4s}}\right]. \quad (12.20)$$

However, this result is only proved in this way for $t\geq 1$, i.e., non-negative values of ζ. This is because the error term on truncating the series inside the square brackets at the mth terms is

$$\{\mathrm{Bi}(-\mu^{\frac{4}{3}}\zeta)+i\,\mathrm{Ai}(-\mu^{\frac{4}{3}}\zeta)\}O(\mu^{-4m})+\{\mathrm{Bi}'(-\mu^{\frac{4}{3}}\zeta)+i\,\mathrm{Ai}'(-\mu^{\frac{4}{3}}\zeta)\}(1+|\zeta|^{\frac{1}{2}})^{-1}O(\mu^{-4m-\frac{2}{3}}).$$

When ζ is negative the terms in Bi and Bi′ cannot be absorbed in the terms in Ai and Ai′ because they are exponentially larger.

That (12.20) *is* valid in the wider range $t\geq -1+\delta$ can be proved as follows. In (8.11) replace μ by $\mu e^{\frac{1}{4}\pi i}$. Then we obtain

$$U\left(\tfrac{1}{2}i\mu^2,e^{\frac{3}{4}\pi i}\mu t\sqrt{2}\right)\sim 2\pi^{\frac{1}{2}}e^{\frac{1}{4}\pi i}\mu^{\frac{1}{3}}g(\mu e^{\frac{1}{4}\pi i})\left(\frac{\zeta}{t^2-1}\right)^{\frac{1}{4}}$$

$$\times\left[\mathrm{Ai}(-\mu^{\frac{4}{3}}\zeta)\sum_{s=0}^{\infty}(-)^s\frac{A_s(\zeta)}{\mu^{4s}}+\frac{\mathrm{Ai}'(-\mu^{\frac{4}{3}}\zeta)}{\mu^{\frac{5}{3}}}\sum_{s=0}^{\infty}(-)^s\frac{B_s(\zeta)}{\mu^{4s}}\right] \quad (12.21)$$

valid when $t\in\mathbf{T}(\frac{3}{4}\pi)$and so including $t\geq -1+\delta$ (cf. fig. 7). Equation (12.1), with t replaced by $-t$, shows that the left-hand side equals

$$2^{-\frac{1}{2}}e^{-\frac{1}{8}\pi\mu^2}e^{-i(\frac{1}{2}\phi_2+\frac{1}{8}\pi)}\left[k^{-\frac{1}{2}}W\left(\tfrac{1}{2}\mu^2,-\mu t\sqrt{2}\right)+ik^{\frac{1}{2}}W\left(\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)\right]. \quad (12.22)$$

Also from (6.1) and (12.7) we derive

$$\begin{aligned}g(\mu e^{\frac{3}{4}\pi i})&\sim 2^{\frac{1}{4}i\mu^2-\frac{1}{4}}e^{\frac{1}{4}i\mu^2}e^{-\frac{3}{8}\pi i+\frac{1}{8}\pi\mu^2}\mu^{-\frac{1}{2}i\mu^2-\frac{1}{2}}(1+iL)^{-1}\\&\simeq e^{\frac{1}{2}\pi\mu^2-\frac{1}{2}\pi i}g(\mu e^{-\frac{1}{4}\pi i})\\&=2^{-\frac{1}{2}}e^{\frac{3}{8}\pi\mu^2}e^{-i(\frac{1}{2}\phi_2+\frac{3}{8}\pi)}l(\mu)\end{aligned} \quad (12.23)$$

(cf. (12.6) and (12.9)). The expansion (12.20) may now be obtained by substituting (12.22) and (12.23) in (12.21), equating real parts and using the first of (12.3).

(iii) *a negative.* When a is negative the differential equation has no turning points on the real axis and we may confine our attention to the exponential-type expansions.

From (2.16), (2.17) and (2.18), we have

$$k^{-\frac{1}{2}}W\left(-\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)+ik^{\frac{1}{2}}W\left(-\tfrac{1}{2}\mu^2,-\mu t\sqrt{2}\right)=2^{\frac{1}{2}}e^{-\frac{1}{8}\pi\mu^2}e^{i(\frac{1}{2}\phi_2+\frac{1}{8}\pi)}U\left(-\tfrac{1}{2}i\mu^2,e^{-\frac{1}{4}\pi i}\mu t\sqrt{2}\right), \quad (12.24)$$

where

$$k=(1+e^{-\pi\mu^2})^{\frac{1}{2}}-e^{-\frac{1}{2}\pi\mu^2}=1+O(e^{-\frac{1}{2}\pi\mu^2}), \quad (12.25)$$

and

$$e^{\frac{1}{2}i\phi_2}=\pm\left\{\Gamma\left(\frac{1}{2}-\frac{1}{2}\,i\mu^2\right)\Big/\left|\Gamma\left(\frac{1}{2}-\frac{1}{2}\,i\mu^2\right)\right|\right\}^{\frac{1}{2}}=e^{\frac{1}{4}i\mu^2}\left(\frac{1}{2}\mu^2\right)^{-\frac{1}{4}i\mu^2}\{1+O(\mu^{-2})\} \tag{12.26}$$

(cf. (12.2)and(12.3)).

In (4.3) replace μ, t by $\mu e^{\frac{1}{4}\pi i}$, $te^{-\frac{1}{2}\pi i}$, respectively. Then we obtain

$$U\left(-\frac{1}{2}\,i\mu^2,e^{-\frac{1}{4}\pi i}\mu t\sqrt{2}\right)\sim g(\mu e^{\frac{1}{4}\pi i})\,\frac{e^{-i\mu^2\xi(-it)}}{(-t^2-1)^{\frac{1}{4}}}\sum_{s=0}^{\infty}(-i)^s\,\frac{u_s(-it)}{(-t^2-1)^{\frac{3}{2}s}}\,\frac{1}{\mu^{2s}}. \tag{12.27}$$

The region of validity of this expansion, $e^{\frac{1}{2}\pi i}\mathbf{S}(\frac{1}{4}\pi)$, includes the whole of the real t-axis; this is clear from figures 2(b), 3(iii) and 1.

Substituting (12.25), (12.26), and (12.27) in (12.24) in the manner of (i) and (ii) and determining the branches of the many-valued functions, we obtain finally

$$W\left(-\tfrac{1}{2}\mu^2,\mu t\sqrt{2}\right)\sim\frac{l(\mu)}{(t^2+1)^{\frac{1}{4}}}\Bigg[\cos\left(\mu^2\bar{\xi}+\tfrac{1}{4}\pi\right)\sum_{s=0}^{\infty}\frac{(-)^s\,\bar{u}_{2s}(t)}{(t^2+1)^{3s}\mu^{4s}}$$

$$-\sin\left(\mu^2\bar{\xi}+\tfrac{1}{4}\pi\right)\sum_{s=0}^{\infty}\frac{(-)^s\bar{u}_{2s+1}(t)}{(t^2+1)^{3s+\frac{3}{2}}\mu^{4s+2}}\Bigg], \tag{12.28}$$

uniformly when $-\infty<t<\infty$. Here $\bar{\xi}\equiv\bar{\xi}(t)$ and $\bar{u}_s(t)$ are given by (11.7), (11.9), and (11.11), and $l(\mu)$ is calculable from (12.11).

The corresponding expansion for the derivative is

$$W'\left(-\tfrac{1}{2}\,\mu^2,\mu t\sqrt{2}\right)\sim-\frac{\mu}{\sqrt{2}}\,l(\mu)(t^2+1)^{\frac{1}{4}}\Bigg[\sin\left(\mu^2\bar{\xi}+\tfrac{1}{4}\pi\right)\sum_{s=0}^{\infty}\frac{(-)^s\bar{v}_{2s}(t)}{(t^2+1)^{3s}\mu^{4s}}$$

$$+\cos\left(\mu^2\bar{\xi}+\tfrac{1}{4}\pi\right)\sum_{s=0}^{\infty}\frac{(-)^s\,\bar{v}_{2s+1}(t)}{(t^2+1)^{3s+\frac{3}{2}}\mu^{4s+2}}\Bigg], \tag{12.29}$$

$\bar{v}_s(t)$ being given by (11.13) and (11.14).

13. Numerical Examples

Here we illustrate the use of the Airy-type expansions by two examples. It may be remarked that applications would be greatly facilitated by the preparation of suitable numerical tables of the functions $g(\mu)$, $l(\mu)$, $\zeta(t)$, $\{\zeta/(t^2-1)\}^{\frac{1}{4}}$ and $A_s(\zeta)$, $B_s(\zeta)$, $C_s(\zeta)$, $D_s(\zeta)$ $(s=0,1,\ldots)$. In the present examples these functions are calculated directly from formulas given in this paper.

It will appear from the examples that the early coefficients in the expansions decrease in magnitude as s increases. This property, shared with uniform asymptotic expansions for Bessel functions of large order ([8], sec. 6), makes the expansions well suited to numerical work.

Example 1. To evaluate $W(a,\pm x)$, $W'(a,\pm x)$ for $a=8$, $x=7$.

We use the expansions (12.18), (12.19), and (12.20). Here

$$\mu=4,\qquad t=7/(4\sqrt{2})=1.23743\ 6867.$$

Hence

$$2^{\frac{1}{2}}e^{\frac{1}{4}\pi\mu^2}=4.05527\ 596\times10^5, \tag{13.1}$$

and from (12.11), (12.12)

$$l(\mu)\sim(2^{\frac{1}{4}}/2)\ (1-0.00000\ 3391-0.00000\ 0006+\ \ldots)=0.59460\ 1538,$$

so that

$$\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}l(\mu)=1.67296\ 798. \tag{13.2}$$

Next, from (7.4) we have

$$\zeta^{\frac{3}{2}}=\frac{3}{4}\,t(t^2-1)^{\frac{1}{2}}-\frac{3}{4}\ln\{t+(t^2-1)^{\frac{1}{2}}\}=0.16932\ 9581.$$

Hence

$$\zeta=0.30607\ 0793,\quad \zeta^{\frac{1}{2}}=0.55323\ 6652,\quad \{\zeta/(t^2-1)\}^{\frac{1}{4}}=0.87122\ 5817. \tag{13.3}$$

From (4.8) and (4.11) we compute

$$\mathscr{A}_0=1,\quad \mathscr{A}_1=-0.59504\ 451,\quad \mathscr{A}_2=2.92479\ 70,\quad \mathscr{A}_3=-26.48264\ 7,$$

and from (8.13)

$$a_0=1,\quad a_1=\frac{5}{48},\qquad a_2=\frac{385}{4608},\qquad a_3=\frac{85085}{6\,63552},$$

$$b_0=1,\quad b_1=-\frac{7}{48},\quad b_2=-\frac{455}{4608},\quad b_3=-\frac{95095}{6\,63552}.$$

Substitution in (8.12) yields

$$A_0(\zeta)=1,\quad A_1(\zeta)=-0.00649\ 0,\quad B_0(\zeta)=-0.03637\ 97,\quad B_1(\zeta)=0.01202.$$

Next, we have $\mu^{\frac{2}{3}}\zeta=1.94342\ 840$, and by interpolation in table VII of [3]

$$\mathrm{Bi}(-\mu^{\frac{2}{3}}\zeta)=-0.39524\ 090,\qquad \mathrm{Bi}'(-\mu^{\frac{2}{3}}\zeta)=0.32386\ 134,$$

$$\mathrm{Ai}(-\mu^{\frac{2}{3}}\zeta)=0.26162\ 602,\qquad \mathrm{Ai}'(-\mu^{\frac{2}{3}}\zeta)=0.59097\ 966.$$

The content of the square brackets on the right of (12.18) is accordingly

$$-0.39524\ 090(1+0.00002\ 535+\ \ldots)+4^{-\frac{2}{3}}\times0.32386\ 134(-0.03637\ 97-0.00004\ 70+\ \ldots)$$
$$=-0.39554\ 353.$$

Substituting this value and (13.1), (13.2), (13.3) in (12.18), we obtain

$$W(8,7)=-\frac{1.67296\,798\times0.87122\,5817\times0.39554\,353}{4.05527\,596\times10^5}=\mathbf{-0.14216\,485\times10^{-5}}.$$

Similarly,

$$W(8,-7)=1.67296\ 798\times0.87122\ 5817\times0.26109\ 871\times4.05527\ 596\times10^5=\mathbf{1.54327\ 57\times10^5}.$$

For the derivatives, we have first

$$2^{-\frac{1}{2}}\pi^{\frac{1}{2}}\mu^{\frac{2}{3}}l(\mu)=1.87784\,306.$$

From (4.14) and (4.16), we compute

$$\mathscr{B}_0=1,\quad \mathscr{B}_1=1.00283\ 725,\quad \mathscr{B}_2=-3.52326\ 59,\quad \mathscr{B}_3=29.91942\ 3,$$

and thence from (8.19)

$$C_0(\zeta)=-0.07833\ 70,\quad C_1(\zeta)=0.00971,\quad D_0(\zeta)=1,\quad D_1(\zeta)=0.00760\ 4.$$

Hence

$$W'(8,7)=-\frac{1.87784\,306\times0.31897\,315}{4.05527\,596\times10^5\times0.87122\,5817}=\mathbf{-0.16953\,613\times10^{-5}},$$

$$W'(8,-7)=\frac{1.87784\,306\times0.59419\,143\times4.05527\,596\times10^5}{0.87122\,5817}=\mathbf{5.19368\,20\times10^5}.$$

The results of this example agree adequately with the entries given in [4], table I.

Example 2. To evaluate the Hermite polynomial $H_n(x)$ for $n=10$, $x=4$ by means of the expansion (8.11).

Here (cf. (11.23))

$$\mu=\sqrt{21}=4.58257\ 569, \qquad t=4/\sqrt{21}=0.87287\ 1561.$$

From (6.1) and (4.12)

$$1/g(\mu)\sim 2^{5.5}e^{5.25}(21)^{-5}(1+0.00198\ 4127-0.00000\ 1052+\ \ldots)=2.11580\ 149\times 10^{-3},$$

and hence

$$2\pi^{\frac{1}{2}}\mu^{\frac{1}{3}}g(\mu)=2.78291\ 471\times 10^{3}.$$

Next, from (11.21)

$$(-\zeta)^{\frac{3}{2}}=\frac{3}{4}\cos^{-1}t-\frac{3}{4}\,t\,(1-t^2)^{\frac{1}{2}}=0.06286\ 6477.$$

Hence

$$\zeta=-0.15810\ 5174, \quad (-\zeta)^{\frac{1}{2}}=0.39762\ 4413, \quad \{\zeta/(t^2-1)\}^{\frac{1}{4}}=0.90271\ 123.$$

When ζ is negative, real forms of equations (8.12) and (4.8) are

$$A_s(\zeta)=(-)^s\sum_{m=0}^{2s} b_m(-\zeta)^{-\frac{1}{2}m}\mathscr{Z}_{2s-m}, \qquad (-\zeta)^{\frac{1}{2}}B_s(\zeta)=(-)^{s-1}\sum_{m=0}^{2s+1} a_m(-\zeta)^{-\frac{1}{2}m}\mathscr{Z}_{2s-m+1}, \quad (13.4)$$

where

$$\mathscr{Z}_s=u_s(t)/(1-t^2)^{\frac{1}{2}s}. \quad (13.5)$$

In the present example we find

$$\mathscr{Z}_0=1, \quad \mathscr{Z}_1=-1.63978\ 319, \quad \mathscr{Z}_2=21.19031\ 25, \quad \mathscr{Z}_3=-516.52363,$$

and thence

$$A_0(\zeta)=1, \quad A_1(\zeta)=-0.01024\ 3, \quad B_0(\zeta)=-0.04317\ 57, \quad B_1(\zeta)=0.0166.$$

Next, $\mu^{\frac{4}{3}}\zeta=-1.20344\ 324$. Interpolation in table I of [3], yields

$$\mathrm{Ai}(\mu^{\frac{4}{3}}\zeta)=0.52582\ 209, \qquad \mathrm{Ai}'(\mu^{\frac{4}{3}}\zeta)=0.10920\ 810.$$

Substituting these results in (8.11), we obtain

$$U(-\tfrac{21}{2},\ 4\sqrt{2})=2.78291\ 471\times 10^{3}\times 0.90271\ 123\times 0.52572\ 856=1.32071\ 865\times 10^{3},$$

and thence using (11.23)

$$H_{10}(4)=2^5e^8U\left(-\frac{21}{2},\ 4\sqrt{2}\right)=\mathbf{1.25984\ 22\times 10^{8}}.$$

The exact value of this polynomial, computed from explicit expressions given on page 75 of [4], is 1259 84224.

Part 5. Zeros and Associated Values

14. Zeros of $U(a, z)$, $U'(a, z)$, $\overline{U}(a, z)$, and $\overline{U}'(a, z)$

(i) *a negative;* $-2(1-\delta)\sqrt{(-a)}\le z\le 2\sqrt{(-a)}$. Let the sth real zeros, counted in descending order away from the point $z=2\sqrt{(-a)}$, of the functions $U(a,z)$, $U'(a,z)$, $\overline{U}(a,z)$, and $\overline{U}'(a,z)$ be denoted by $u_{a,s}$, $u'_{a,s}$, $\bar{u}_{a,s}$, and $\bar{u}'_{a,s}$, respectively.

The expansion (8.11) is of exactly the same form as the expansion (4.24) of [8]. The analysis of section 7 of that paper is therefore immediately applicable, and we find

$$u_{a,s}\sim 2^{\frac{1}{2}}\mu\left[p_0(\alpha)+\frac{p_1(\alpha)}{\mu^4}+\frac{p_2(\alpha)}{\mu^8}+\ \ldots\right] \quad (14.1)$$

as $\mu \equiv \sqrt{(-2a)} \to +\infty$, s fixed. Here

$$\alpha = \mu^{-\frac{2}{3}} a_s, \tag{14.2}$$

a_s denoting the sth negative zero of the function Ai and not to be confused with the a_m of (8.13). The coefficients $p_r(\zeta)$ are given by

$$p_0(\zeta) = t(\zeta), \qquad p_1(\zeta) = \alpha_1(\zeta) t'(\zeta), \qquad p_2(\zeta) = \alpha_2(\zeta) t'(\zeta) + \frac{1}{2} \{\alpha_1(\zeta)\}^2 t''(\zeta), \ldots, \tag{14.3}$$

where $t(\zeta)$ is the function inverse to $\zeta(t)$, defined by (11.21), and the $\alpha_r(\zeta)$ are given by relations of the form of (7.13) of [8]. Explicitly,

$$p_1(\zeta) = \frac{t^3 - 6t}{24(t^2-1)^2} + \frac{5}{48\{(t^2-1)\zeta^3\}^{\frac{1}{2}}}. \tag{14.4}$$

The corresponding expansion for the associated value $U'(a, u_{a,s})$ can be found by substituting (14.1) in (8.15). More convenient expressions for the coefficients are obtained however, if we substitute instead in the equation

$$U'(a, u_{a,s}) = \pm (2\pi)^{\frac{1}{2}} \left\{ \Gamma\left(\frac{1}{2} - a\right) \right\}^{\frac{1}{2}} \left(-\frac{du_{a,s}}{ds} \right)^{-\frac{1}{2}} \tag{14.5}$$

obtained from eq (2.15) of [6] and the second of (2.13) in the present paper. This yields the required result in the form

$$U'(a, u_{a,s}) \sim \pi^{\frac{1}{4}} \mu^{\frac{1}{6}} \left\{ \Gamma\left(\frac{1}{2} + \frac{1}{2}\mu^2\right) \right\}^{\frac{1}{2}} \mathrm{Ai}'(a_s) \psi(\alpha) \left[1 + \sum_{r=1}^{\infty} \frac{P_r(\alpha)}{\mu^{4r}} \right], \tag{14.6}$$

where

$$\psi(\zeta) \equiv \frac{1}{\phi(\zeta)} = \left(\frac{t^2-1}{\zeta} \right)^{\frac{1}{4}}, \tag{14.7}$$

and the coefficients $P_r(\zeta)$ satisfy the asymptotic identity

$$\left[1 + \sum_{r=1}^{\infty} \frac{P_r(\zeta)}{\mu^{4r}} \right]^{-2} \sim 1 + \{\psi(\zeta)\}^2 \sum_{r=1}^{\infty} \frac{p_r'(\zeta)}{\mu^{4r}}. \tag{14.8}$$

Thus

$$P_1(\zeta) = -\frac{1}{2} \{\psi(\zeta)\}^2 p_1'(\zeta), \qquad P_2(\zeta) = \frac{3}{2} \{P_1(\zeta)\}^2 - \frac{1}{2} \{\psi(\zeta)\}^2 p_2'(\zeta). \tag{14.9}$$

Explicitly,

$$P_1(\zeta) = \frac{t^4 - 15t^2 - 6}{48(t^2-1)^3} + \frac{5t}{96(t^2-1)^{\frac{5}{2}} \zeta^{\frac{3}{2}}} + \frac{5}{64\zeta^3}. \tag{14.10}$$

For the zeros of the derivatives of the parabolic cylinder functions, we may show by similar analysis that

$$u'_{a,s} \sim 2^{\frac{1}{3}} \mu \left[q_0(\beta) + \frac{q_1(\beta)}{\mu^4} + \frac{q_2(\beta)}{\mu^8} + \ldots \right], \tag{14.11}$$

where

$$\beta = \mu^{-\frac{2}{3}} a'_s, \tag{14.12}$$

$$q_0(\zeta) = t(\zeta), \qquad q_1(\zeta) = \beta_1(\zeta) t'(\zeta), \qquad q_2(\zeta) = \beta_2(\zeta) t'(\zeta) + \frac{1}{2} \{\beta_1(\zeta)\}^2 t''(\zeta), \ldots, \tag{14.13}$$

and the $\beta_r(\zeta)$ are given by eq (7.22) of [8].

The equation corresponding to (14.5) is

$$U(a, u'_{a,s}) = \pm (2\pi)^{\frac{1}{2}} \left\{ \Gamma\left(\frac{1}{2} - a\right) \right\}^{\frac{1}{2}} \left(-a - \frac{1}{4} u'^2_{a,s} \right)^{-\frac{1}{2}} \left(-\frac{du'_{a,s}}{ds} \right)^{-\frac{1}{2}}, \tag{14.14}$$

and from it we derive

$$U(a,u'_{a,s})\sim\frac{(4\pi)^{\frac{1}{4}}}{\mu^{\frac{1}{6}}}\left\{\Gamma\left(\frac{1}{2}+\frac{1}{2}\mu^2\right)\right\}^{\frac{1}{2}}\mathrm{Ai}\,(a'_s)\phi(\beta)\left[1+\sum_{r=1}^{\infty}\frac{Q_r(\beta)}{\mu^{4r}}\right], \tag{14.15}$$

in which the coefficients satisfy the identity

$$\left[1+\sum_{r=1}^{\infty}\frac{Q_r(\zeta)}{\mu^{4r}}\right]^{-2}\sim\frac{\{\phi(\zeta)\}^2}{\zeta}\left[\left\{\sum_{r=0}^{\infty}\frac{q_r(\zeta)}{\mu^{4r}}\right\}^2-1\right]\left[\sum_{r=0}^{\infty}\frac{q'_r(\zeta)}{\mu^{4r}}\right]. \tag{14.16}$$

For the function $\overline{U}(a,z)$, the expansions corresponding to (14.1), (14.6), (14.11), and (14.15) are obtained on replacing the symbols U, u, Ai, and a_s by $\overline{U}$, $\bar{u}$, Bi, and b_s, respectively.

(ii) *a negative; other zeros.* Since the zeros of the function Ai are all real and negative, it follows from (8.11) that for all sufficiently large positive values of the parameter μ, the zeros of $U(-\frac{1}{2}\mu^2,\,\mu t\sqrt{2})$ in the t-domain $\mathbf{T}(0)$ (the unshaded region of fig. 7(a)), all lie in the interval $-1<t<1$. This is also true of the zeros of the derived function $U'(-\frac{1}{2}\mu^2,\mu t\sqrt{2})$. The function $\overline{U}(-\frac{1}{2}\mu^2,\mu t\sqrt{2})$ and its derivative however, have complex zeros in this region corresponding to the complex zeros of Bi and Bi′ (cf. [8], appendix).

In the complementary region $C\mathbf{T}(0)$ the distribution of the zeros of $U(-\frac{1}{2}\mu^2,\mu t\sqrt{2})$ can be investigated by reversion of the expansion (9.7). The situation is more complicated; the pattern of the zeros depends largely upon the non-integer part of $\frac{1}{2}\mu^2$. It should be noted that expansions of zeros in the interval $-1<t<1$ of fixed enumeration, counted away from $t=-1$, come within this category.

(iii) *a complex.* The extension of the results of this section to general values of a presents no difficulty. We readily see, for example, that when $\epsilon\leq\arg\mu\leq\pi-\epsilon$ the zeros of $U(-\frac{1}{2}\mu^2,\,\mu t\sqrt{2})$ in the t-domain $\mathbf{T}(\arg\mu)$ are asymptotically close to the principal curve AQ, illustrated in figure 4.

15. Zeros of $W(a,x)$ and $W'(a,x)$

We denote the sth positive zeros of the functions $W(a,x)$, $W'(a,x)$, $W(a,-x)$ and $W'(a,-x)$, by $w_{a,s}$, $w'_{a,s}$, $\overline{w}_{a,s}$ and $\overline{w}'_{a,s}$, respectively.

(i) *a positive.* Writing $\mu=\sqrt{(2a)}$ and

$$\alpha=\mu^{-\frac{2}{3}}b_s,\qquad\beta=\mu^{-\frac{2}{3}}b'_s, \tag{15.1}$$

where b_s and b'_s are the sth negative zeros of Bi and Bi′ respectively, we find by reversion of (12.18) and (12.20)

$$w_{a,s}\sim 2^{\frac{1}{2}}\mu\left[p_0(-\alpha)-\frac{p_1(-\alpha)}{\mu^4}+\frac{p_2(-\alpha)}{\mu^8}-\ \ldots\right], \tag{15.2}$$

$$2^{\frac{1}{2}}e^{\frac{1}{2}\pi a}W'(a,w_{a,s})\sim-2^{-\frac{1}{4}}\pi^{\frac{1}{2}}\mu^{\frac{1}{6}}\mathrm{Bi}'(b_s)\psi(-\alpha)\left[1+\sum_{r=1}^{\infty}(-)^r\frac{P_r(-\alpha)}{\mu^{4r}}\right], \tag{15.3}$$

$$w'_{a,s}\sim 2^{\frac{1}{2}}\mu\left[q_0(-\beta)-\frac{q_1(-\beta)}{\mu^4}+\frac{q_2(-\beta)}{\mu^8}-\ \ldots\right], \tag{15.4}$$

$$2^{\frac{1}{2}}e^{\frac{1}{2}\pi a}W(a,w'_{a,s})\sim 2^{\frac{1}{4}}\pi^{\frac{1}{2}}\mu^{-\frac{1}{6}}\,\mathrm{Bi}\,(b'_s)\phi(-\beta)\left[1+\sum_{r=1}^{\infty}(-)^r\frac{Q_r(-\beta)}{\mu^{4r}}\right], \tag{15.5}$$

as $a\to+\infty$, uniformly with respect to s. Here $\psi(\zeta)$, $\phi(\zeta)$, $p_r(\zeta)$, $P_r(\zeta)$, $q_r(\zeta)$ and $Q_r(\zeta)$ are the functions defined in section 14(i).

With α, β given by (14.2) and (14.12) instead of (15.1) and Bi, b replaced by Ai, a, the right-hand sides of (15.2), (15.3), (15.4), and (15.5) are the corresponding expansions of $\overline{w}_{a,s}$, $-2^{-\frac{1}{2}}e^{-\frac{1}{2}\pi a}W'(a,-\overline{w}_{a,s})$, $\overline{w}'_{a,s}$ and $2^{-\frac{1}{2}}e^{-\frac{1}{2}\pi a}W(a,-\overline{w}'_{a,s})$, respectively.

(ii) *a negative.* Let $\mu=\sqrt{(-2a)}$, and

$$t=\tau(\bar{\xi}) \tag{15.6}$$

be the function inverse to $\bar{\xi}(t)$, defined by (11.7). Then by reversion of the expansion (12.28), we find

$$w_{a,s}\sim 2^{\frac{1}{2}}\mu\left[\tau+\frac{e_1(\tau)}{\mu^4}+\frac{e_2(\tau)}{\mu^8}+\cdots\right] \tag{15.7}$$

as $a\to-\infty$, uniformly with respect to s. Here

$$\tau=\tau\left\{\frac{(s-\frac{3}{4})\pi}{\mu^2}\right\}, \tag{15.8}$$

and the leading coefficients are given by the formulas

$$e_1(\tau)=\frac{\tau^3+6\tau}{24(\tau^2+1)^2},\qquad e_2(\tau)=\frac{56\tau^9+247\tau^7+291\tau^5-2520\tau^3+3780\tau}{5760(\tau^2+1)^5}. \tag{15.9}$$

The expansion of the associated value of the derivative can be obtained by substitution of (15.7) in the equation

$$W'(a,w_{a,s})=\pm\pi^{\frac{1}{2}}\left(\frac{dw_{a,s}}{ds}\right)^{-\frac{1}{2}} \tag{15.10}$$

obtained from eq (2.15) of [6] and (2.19). Thence we derive

$$W'(a,w_{a,s})\sim(-)^s2^{-\frac{1}{4}}\mu^{\frac{1}{2}}(\tau^2+1)^{\frac{1}{4}}\left[1+\sum_{r=1}^{\infty}\frac{E_r(\tau)}{\mu^{4r}}\right], \tag{15.11}$$

where

$$E_1(\tau)=\frac{\tau^4+15\tau^2-6}{48(\tau^2+1)^3},\qquad E_2(\tau)=\frac{112\tau^{10}+489\tau^8-98\tau^6-34995\tau^4+80460\tau^2-7020}{23040(\tau^2+1)^6}. \tag{15.12}$$

In a similar manner we obtain from (12.29)

$$w'_{a,s}\sim 2^{\frac{1}{2}}\mu\left[\tau+\frac{f_1(\tau)}{\mu^4}+\frac{f_2(\tau)}{\mu^8}+\cdots\right], \tag{15.13}$$

where now

$$\tau=\tau\left\{\frac{(s-\frac{1}{4})\pi}{\mu^2}\right\}, \tag{15.14}$$

and

$$f_1(\tau)=\frac{\tau^3-6\tau}{24(\tau^2+1)^2},\qquad f_2(\tau)=\frac{56\tau^9+247\tau^7+651\tau^5+2640\tau^3-3420\tau}{5760(\tau^2+1)^5}. \tag{15.15}$$

Finally, from

$$W(a,w'_{a,s})=\pm\pi^{\frac{1}{2}}\left(\frac{1}{4}w'^2_{a,s}-a\right)^{-\frac{1}{2}}\left(\frac{dw'_{a,s}}{ds}\right)^{-\frac{1}{2}}, \tag{15.16}$$

we find

$$W(a,w'_{a,s})\sim(-)^s2^{\frac{1}{4}}\mu^{-\frac{1}{2}}(\tau^2+1)^{-\frac{1}{4}}\left[1+\sum_{r=1}^{\infty}\frac{F_r(\tau)}{\mu^{4r}}\right], \tag{15.17}$$

where τ is again given by (15.14), and

$$F_1(\tau)=\frac{-\tau^4-9\tau^2+6}{48(\tau^2+1)^3},\qquad F_2(\tau)=\frac{-112\tau^{10}-479\tau^8-142\tau^6+25725\tau^4-67500\tau^2+7380}{23040(\tau^2+1)^6}. \tag{15.18}$$

If in the right-hand sides of (15.7) and (15.11) we take the value (15.14) for τ instead of (15.8), we obtain the asymptotic expansions of $\overline{w}_{a,s}$ and $-W'(a,-\overline{w}_{a,s})$, respectively. Similarly the right-hand sides of (15.13) and (15.17) with τ given by (15.8) are the expansions of $\overline{w}'_{a,s}$ and $-W(a,-\overline{w}'_{a,s})$.

It may be noticed that some of the numerical coefficients occurring in the expressions (14.4) and (14.10) for the coefficients in the reverted Airy-type expansions are the same as those in the expressions (15.9) and (15.12) pertaining to the reverted exponential-type expansions. This is no mere coincidence; we may expect that by suitable re-expansion of the reverted Airy forms, and subsequent comparison with exponential-type expansions obtained by reversion of (5.11), we shall obtain explicit expressions for the coefficients p_r and P_r in terms of the ϵ_r and E_r respectively, analogous to the formulas (8.12). Similarly for q_r and Q_r.

16. Numerical Examples

Example 1. To compute the smallest positive zero of $W(10,x)$ and the corresponding $W'(10,x)$.

We use (15.2) and (15.3) with $\mu=\sqrt{20}=4.47213\,595$, $s=1$. Entering table V of [3], we find

$$b_1=-1.17371\,322, \qquad \mathrm{Bi}'(b_1)=0.60195\,789.$$

Hence

$$\zeta=-\alpha=-\mu^{-\frac{2}{3}}b_1=0.15929\,7392.$$

The value of t is given by

$$t(t^2-1)^{\frac{1}{2}}-\ln\{t+(t^2-1)^{\frac{1}{2}}\}=\frac{4}{3}\zeta^{\frac{3}{2}}=0.08477\,1864.$$

Solving this equation by successive approximation, using Newton's rule, we find

$$t=1.12489\,6229=p_0(-\alpha),$$

and thence from (14.4)

$$p_1(-\alpha)=0.02960\,04.$$

Hence

$$w_{10,1}\sim 2^{\frac{1}{2}}\times 4.47213\,595(1.12489\,6229-0.00007\,4001+\ \ldots)=\mathbf{7.11400\,040}.$$

For the derivative, we compute from (14.7) and (14.10)

$$1/\psi(-\alpha)=\phi(-\alpha)=0.88019\,8053, \qquad P_1(-\alpha)=0.00942\,2.$$

Substitution in (15.3) then yields

$$W'(10,w_{10,1})\sim-\frac{2^{-\frac{1}{4}}\pi^{\frac{1}{2}}(20)^{\frac{1}{12}}\times 0.60195\,789}{e^{5\pi}\times 0.88019\,8053}(1-0.00002\,356+\ \ldots)=\mathbf{-1.39416\,865\times 10^{-7}}.$$

By inverse and direct interpolation in table III of [4], we find

$$w_{10,1}=7.11400\,000, \qquad W'(10,w_{10,1})=-1.39416\,86\times 10^{-7}.$$

The discrepancy in the values of $w_{10,1}$ is undoubtedly due to neglect of the third term of the asymptotic series.

Example 2. To compute the fifth positive zero of $W'(-10,x)$ and the corresponding $W(-10,x)$.

We use (15.13) and (15.17). From (15.14), (15.6), and (11.7) we see that the appropriate value of τ satisfies

$$\tau(\tau^2+1)^{\frac{1}{2}}+\ln\{\tau+(\tau^2+1)^{\frac{1}{2}}\}=2\bar{\xi}=2(5-\tfrac{1}{4})\pi/20=1.49225\,6510.$$

Solving by successive approximation, we find

$$\tau=0.69390\,0720, \qquad (\tau^2+1)^{\frac{1}{2}}=1.21716\,8110,$$

and thence from (15.15)

$$f_1(\tau)=-0.07269\ 508, \qquad f_2(\tau)=-0.03321.$$

Thus

$$w'_{-10,5}\sim 2^{\frac{1}{2}}\times 4.47213\ 595(0.69390\ 0720-0.00018\ 1738-0.00000\ 0208+\ \ldots)=\mathbf{4.38746\ 276}.$$

From (15.18) we compute

$$F_1(\tau)=0.00919\ 198, \qquad F_2(\tau)=-0.07882,$$

and thence

$$W(-10,w'_{-10,5})\sim -2^{\frac{1}{4}}(20)^{-\frac{1}{4}}(1.21716\ 8110)^{-\frac{1}{2}}(1+0.00002\ 2980-0.00000\ 0493+\ \ldots)$$
$$=\mathbf{-0.50972\ 3297}.$$

The results of this example agree with values obtained by interpolation in table III of [4].

References

[1] C. G. Darwin, On Weber's function, Quart. J. Mech. Appl. Math. **2,** 311 (1949).
[2] A. Erdélyi, M. Kennedy, and J. L. McGregor, Parabolic cylinder functions of large order, J. Rat. Mech. Anal. **3,** 459 (1954).
[3] J. C. P. Miller, The Airy integral, British Assoc. Math. Tables, Part-Volume B (Cambridge University Press, Cambridge, Eng. 1946).
[4] J. C. P. Miller, Tables of Weber parabolic cylinder functions, London: H. M. Stationery Office (1955).
[5] L. M. Milne-Thomson, The calculus of finite differences, London: Macmillan (1933).
[6] F. W. J. Olver, A new method for the evaluation of zeros of Bessel functions and of other solutions of second-order differential equations, Proc. Cambridge Philos. Soc. **46,** 570 (1950).
[7] F. W. J. Olver, The asymptotic solution of linear differential equations of the second order for large values of a parameter, Phil. Trans. Roy. Soc. London, Ser. A **247,** 307 (1954).
[8] F. W. J. Olver, The asymptotic expansion of Bessel functions of large order, Phil. Trans. Roy. Soc. London, Ser. A **247,** 328 (1954).
[9] F. W. J. Olver, Uniform asymptotic expansions of solutions of linear second-order differential equations for large values of a parameter, Phil. Trans. Roy. Soc. London, Ser. A **250,** 479 (1958).
[10] N. Schwid, The asymptotic forms of the Hermite and Weber functions, Trans. Amer. Math. Soc. **37,** 339 (1935).
[11] G. N. Watson, A theory of asymptotic series, Phil. Trans. Roy. Soc. London, Ser. A **211,** 279 (1911).
[12] G. N. Watson, The harmonic functions associated with the parabolic cylinder, Proc. London Math. Soc. **17,** 116 (1918).
[13] E. T. Whittaker, On the functions associated with the parabolic cylinder in harmonic analysis, Proc. London Math. Soc. **35,** 417 (1903).
[14] E. T. Whittaker and G. N. Watson, A course of modern analysis (Cambridge University Press, Cambridge, Eng., 1952).
[15] N. D. Kazarinoff, Asymptotic theory of second order differential equations with two simple turning points, Arch. Rat. Mech. Anal. **2,** 129 (1958).

Washington, D. C.

(Paper 63B2-14)

519835—59——7

Publications of the National Bureau of Standards

(including papers in outside journals)

Selected Abstracts

Diffraction of electromagnetic waves by smooth obstacles for grazing angles, J. R. Wait and A. M. Conda, *J. Research NBS* **63D,** *No. 2, 18 (1959).*

The diffraction of electromagnetic waves by a convex cylindrical surface is considered. Attention is confined primarily to the region near the light-shadow boundary. The complex-integral representation for the field is utilized to obtain a correction to the Kirchoff theory. Numerical results are presented which illustrate the influence of surface curvature and polarization on the diffraction pattern. Good agreement with the experimental results of Bachynski and Neugebauer is obtained. The effect of finite conductivity is also considered.

Path antenna gain in an exponential atmosphere, W. J. Hartman and R. E. Wilkerson, *J. Research NBS* **63D,** *No. 3, 273 (1959).*

The problem of determining path antenna gain is treated here in greater detail than previously. The method used here takes into account for the first time the exponential decrease of the gradient of refractive index with height, and a scattering cross section inversely proportional to the fifth power of the scattering angle. Results are given for all combinations of beamwidths and path geometry, assuming that symmetrical beams are used on both ends of the path and that atmospheric turbulence is isotropic. The result appears as a function of both of the beamwidths, in addition to other parameters, and thus the loss in gain cannot be determined independently for the transmitting and receiving antennas. The values of the loss in gain are generally lower than the previous estimates for which a comparison is possible.

Pattern synthesis for slotted-cylinder antennas, J. R. Wait and James Householder, *J. Research NBS* **63D,** *No. 3, 303 (1959).*

The radiation from a cylinder excited by an array of axial slots is discussed. A procedure for synthesizing a given radiation pattern is developed with particular attention being paid to a Tchebyscheff type pattern. Specifying the side lobe level and the width of the main beam, the required source distributions are computed for a number of cases. The effect of using a finite number of slot elements to approximate the continuous source distribution is also considered.

Convexity of the field of a linear transformation, A. J. Goldman and M. Marcus, *Can. Math. Bul.* **2,** *No. 1, 15 (1959).*

The field $F(A)$ of a linear transformation A of unitary n-space U_n is the set of complex numbers $z=(Ax,x)$, where x ranges over the unit sphere in U_n. The convexity of $F(A)$ (as a subset of the z-plane) is proved by a simple inductive argument which reduces the essential computations to the case $n=2$.

Graphical diagnosis of interlaboratory test results, W. J. Youden, *Ind. Qual. Control* **XV,** *No. 11, 1 (1959).*

This analysis of interlaboratory test results depends upon the availability from a number of laboratories of a single test result for each of two materials. The two results from each laboratory are used to plot a point using the x-axis for one material and the y-axis for the other. The resulting configuration of the points from the several laboratories permits deductions on the prevalence and extent of laboratory bias, the presence of sampling variation and the occurrence of blunders. In addition the graph provides an estimate of the precision of the test procedure results.

The calculation of the field in a homogeneous conductor with a wavy interface, J. R. Wait, *Proc. IRE* **47,** *1155 (1959).*

The field at any depth in a homogeneous conductor with a wavy interface is calculated. For purposes of illustration, the field above the interface is taken to be a uniform plane wave traveling in the horizontal direction.

Lower bounds for eigenvalues with application to the helium atom, N. W. Bazley, *Proc. Nat. Acad. Sci. U.S.* **45,** *No. 6, 850 (1959).*

Let A be a self-adjoint operator with domain D in a Hilbert space. Suppose $A=A'+\overline{A}$ where $\overline{A}$ is self-adjoint and A' is positive definite. The eigenvalue problem for $\overline{A}$, whose solution we assume known, gives rough lower bounds. If $\overline{u}_i$ $(i=1, \ldots, k)$ are k discrete eigenvectors of $\overline{A}$ and if $p_i=(a')^{-1}\overline{u}_i (i=1, \ldots, k)$ exist then one can substantially improve the lower bounds. The theory is applied to helium atom operator.

List of Titles *

Journal of Research, Section A. Physics and Chemistry, Volume 63A, No. 2, September–October 1959. 70 cents.

Reflection of fast neutrons from water. Martin J. Berger and John W. Cooper.

Some vibrational-rotational bands of deuterated methanes. Harry C. Allen, Jr., and Earle K. Plyler.

Precise coulometric titration of acids and bases. John K. Taylor and Stanley W. Smith.

Heat of formation of titanium tetraiodide. W. H. Johnson, A. A. Gilliland, and E. J. Prosen.

Reactions of pentafluorohalobenzenes. Walter J. Pummer and Leo A. Wall.

Tritium-labeled compounds I. Radioassay of tritium-labeled compounds in "infinitely thick" films with a windowless, gas-flow, proportional contour. Horace S. Isbell, Harriet L. Frush, and Ruth Peterson.

Tritium-labeled compounds II. General-purpose apparatus, and procedures for the preparation, analysis, and use of tritium oxide and tritium-labeled lithium borohydride. Horace S. Isbell and Joseph D. Moyer.

Preparation of trichloride and tetrachloride of molybdenum. Dwight E. Couch and Abner Brenner.

Infrared spectra of asphalts. Burton D. Beitchman.

Journal of Research, Section C. Engineering and Instrumentation, Volume 63C, No. 2, October–December 1959. 75 cents.

Conical coaxial capacitors and their advantages. M. C. Selby.

A photoelectric followup and recording system, and its application to remote observations of the beam in high precision balances. H. A. Bowman and L. B. Macurdy.

A stroboscopic vibration analyzer. S. Edelman, R. Brooks, S. Saito, E. Jones, and E. R. Smith.

Evaluation of lens distortion by the inverse nodal slide. Francis H. Washer and Walter R. Darling.

Evaluation of lens distortion by the modified goniometric method. Francis E. Washer and Walter R. Darling.

Proposed criteria for defining load failure of beams, floors, and roof constructions during fire tests. J. V. Ryan and A. F. Robertson.

Conductive flooring for hospital operating rooms. Thomas H. Boone, Francis L. Hermach, Edgar H. MacArthur, and Rita C. McAuliff.

Measurement of the aging of rubber vulcanizates. J. Mandel, F. L. Roth, M. N. Steel, and R. D. Stiehler.

Journal of Research, Section D. Radio Propagation, Volume 63D, No. 2, September-October 1959. 70 cents.

Stratification in the lower ionosphere. C. Ellyett and J. M. Watts.

Effect of small irregularities on the constitutive relations for the ionosphere. K. G. Budden.

Ionospheric investigations using the sweep-frequency pulse technique at oblique incidence. Vaughn Agy and Kenneth Davis.

Fields in electrically short ground systems: an experimental study. A. N. Smith and T. E. Devaney.

Diffraction of electromagnetic waves by smooth obstacles for grazing angles. James R. Wait and Alyce M. Conda.

Very-low-frequency radiation spectra of lightning discharges. W. L. Taylor and A. G. Jean.

Radio-wave scattering by tropospheric irregularities. Albert D. Wheelon.

Study at 1046 megacycles per second of the reflection coefficient of irregular terrain at grazing angles. Raymond E. McGavin and Leo J. Maloney.

Synoptic study of the vertical distribution of the radio refractive index. B. R. Bean, L. P. Riggs, and J. D. Horn.

Journal of Research, Section D. Radio Propagation, Volume 63D, No. 3, November-December 1959. 70 cents.

Radio-refractive-index climate near the ground. B. R. Bean and J. D. Horn.

Path antenna gain in an exponential atmosphere. W. J. Hartman and R. E. Wilkerson (see abstracts above).

Effect of atmospheric horizontal inhomogeneity upon ray tracing. B. R. Bean and B. A. Cahoon.

Correlation of solar noise fluctuations in harmonically related bands. L. R. O. Storey.

A monochromatic low-latitude aurora. F. E. Roach and E. Marovich.

Pattern synthesis for slotted-cylinder antennas. James R. Wait and James Householder (see abstracts above).

Central Radio Propagation Laboratory exponential reference atmosphere. B. R. Bean and G. D. Thayer.

Excitation mechanisms of the oxygen 5577 emission in the upper atmosphere. E. Tandberg-Hanssen and F. E. Roach.

Method for measuring local electron density from an artificial satellite. L. R. O. Storey.

Thermodynamic properties of helium at low temperatures and high pressures, D. B. Mann and R. B. Stewart, TN8 (PB151367) $1.25.

Transmission loss in radio propagation—II, K. A. Norton, TN12 (PB151371) $3.00.

Technical considerations leading to an optimum allocation of radio frequencies in the band 25 to 60 Mc, K. A. Norton, TN13 (PB151372) $2.50.

The role of government research laboratories, A. V. Astin, Elec. Engr. **78,** No. 7, 738 (1959).

The ampere, F. B. Silsbee, Proc. IRE **47,** No. 5, 643 (1959).

On the theory of reflection from a wire grid parallel to an interface between homogeneous media (II), J. R. Wait, Appl. Sci. Research **7,** 355 (1959).

The basis of our measuring system, A. G. McNish, Proc. IRE **47,** No. 5, 636 (1959).

Reflectors for a microwave Fabry-Perot interferometer, W. Culshaw, IRE Trans. on Microwave Theory and Techniques, Vol. **MTT-7,** No. 2, 221 (1959).

The research frontier, I. C. Schoonover, Saturday Rev. p. 58 (May 2, 1959).

The nature, cause and effect of the porosity in electrodeposits, III. Microscopic detection of porosity, F. Ogburn and D. W. Ernst, Plating **46,** 831 (1959).

Paper test data from pendulum and inertialess testers, R. J. Capott, S. B. Newman, and J. Mandel, TAPPI **42,** No. 6, 480 (1959).

Factors affecting the accuracy of distortion measurements made on the nodal slide optical bench, F. E. Washer and W. R. Darling, J. Opt. Soc. Am. **49,** No. 6, 517 (1959).

Mechanism of contraction in the muscle fiber-ATP system, L. Mandelkern, A. S. Posner, A. F. Diorio, and K. Laki, Proc. Nat. Acad. Sci. **45,** No. 6, 814 (1959).

Microwave spectrum of methyl germane, V. W. Laurie, J. Chem. Phys. **30,** No. 5, 1210 (1959).

IGY instruction manual, Part 1: World days and communications, A. H. Shapley, Annals of International Geophys. Year **7,** Pt. 1, 1 (Pergamon Press, New York, N.Y., 1959).

Measurement of ozone in terms of its optical absorption, R. Stair, Advances in Chem. Series of the Am. Chem. Soc., No. 21, 269 (1959).

On the perturbation of the vibrational equilibrium distribution of reactant molecules by chemical reactions, K. E. Shuler, 7th Symp. (International) on Combustion, London and Oxford, August 28 to September 3, 1958, Combustion Institute, p. 87 (Butterworths Sci. Pub., London, England, 1958).

Rack for standard resistors, P. H. Lowrie, Jr., Rev. Sci. Instr. **30,** No. 4, 291 (1959).

Spectroscopic evidence for triatomic nitrogen in solids at very low temperature, M. Peyron, E. M. Horl, H. W. Brown, and H. P. Broida, J. Chem. Phys. **30,** No. 5, 1304 (1959).

The evaluation of small color differences: Part I. Visual observations, J. C. Richmond and W. N. Harrison, Am. Ceram. Bul. **38,** No. 6, 292 (1959).

**Publications for which a price is indicated (except for Technical Notes) are available only from the Superintendent of Documents, U.S. Government Printing Office, Washington 25, D.C. (foreign postage, one-fourth additional). Technical Notes are available only from the Office of Technical Services, U.S. Department of Commerce, Washington 25, D.C. (Order by PB number.) Reprints from outside journals and the NBS Journal of Research may often be obtained directly from the authors.*

INDEX TO VOLUME 63B

July to December 1959

A

Page

Analytic comparison of suggested configurations for automatic mail sorting equipment ... 83
Analytical integration of the differential equation for water storage ... 43
Applications of a theorem on partitioned matrices ... 73
Astronomy, new method of solution for unretarded satellite orbits ... 105

B

Bender, B. K., A. J. Goldman, Analytic comparison of suggested configurations for automatic mail sorting equipment ... 83
Boundary layers, compressible turbulent, with heat transfer and pressure gradient in flow direction ... 53

C

Capacity requirement of a mail sorting device: II ... 79
Combinatorial analysis of a mail sorting device ... 79
Compressible turbulent boundary layers with heat transfer and pressure gradient in flow direction ... 53
Configurations for automatic mail sorting equipment, analytic comparison ... 83

E

Effect of sudden water release on the reservoir free outflow hydrograph ... 117
Equation, differential, analytical integration for water storage ... 43
Expansions, uniform asymptotic, for Weber parabolic cylinder functions of large orders ... 131

G

Goldberg, K., Principal submatrices of a full-rowed non-negative matrix ... 19
Goldman, A. J., Capacity requirement of a mail sorting device: II ... 79
—, Zeros of certain polynomials ... 21
—, *B. K. Bender,* Analytic comparison of suggested configurations for automatic mail sorting equipment ... 83
Greub, W., Relations between summation methods and integral transformations ... 1

H

Hall, W. J., W. M. Rogers, R. L. Powell, Tables of transport integrals: a supplement ... 23
Haynsworth, E. V., Applications of a theorem on partitioned matrices ... 73
Hydrograph, sudden water release ... 117
Hydrographs, river flow ... 43

I

Integral transformations, relations between, and summation methods ... 1
Integrals, tables of transport: a supplement ... 23

L

Layers, compressible turbulent boundary, with heat transfer and pressure gradient in flow direction ... 53
Lemma, Watson's modification ... 15
Lens design: a new approach ... 31
Lototsky method for summation ... 1

M

Mail sorting device, capacity requirement ... 79
Mail sorting equipment, analytic comparison of suggested configurations ... 83
Matrices, partitioned, applications of a theorem ... 73
Matrix, principal submatrices of a full-rowed non-negative ... 19

N

New method of solution for unretarded satellite orbits ... 105
Numerical analysis:
Applications of a theorem on partitioned matrices ... 73
Matrix, principal submatrices of a full-rowed non-negative ... 19
Summation methods and integral transformations ... 1
Uniform asymptotic expansions for Weber parabolic cylinder functions of large orders ... 131
Zeros of certain polynomials ... 21

O

Page

Oberhettinger, F., On a modification of Watson's lemma ... 15
Olver, F. W. J., Uniform asymptotic expansions for Weber parabolic cylinder functions of large orders ... 131
On a modification of Watson's lemma ... 15
Orbits, satellite, unretarded, new method of solution ... 105

P

Polynomials, zeros ... 21
Powell, R. L., W. M. Rogers, W. J. Hall, Tables of transport integrals: a supplement ... 23
Principal submatrices of a full-rowed non-negative matrix ... 19

R

Relations between summation methods and integral transformations ... 1
Rogers, W. M., W. J. Hall, R. I. Powell, Tables of transport integrals: a supplement ... 23

S

Satellite orbits, unretarded, new method of solution ... 105
Sorting device (mail), capacity requirement ... 79
Sorting equipment (mail), analytic comparison of suggested configurations ... 83
Starroudis, O. N., Lens design: a new approach ... 31
Submatrices of a full-rowed non-negative matrix, principal ... 19
Summation methods and integral transformations, relations between ... 1

T

Tables of transport integrals: a supplement ... 23
Transformations, integral, relations between and summation methods ... 1
Transport integrals, tables: a supplement ... 23
Turbulent boundary layers with heat transfer and pressure gradient in flow direction ... 53

U

Uniform asymptotic expansions for Weber parabolic cylinder functions of large orders ... 131

V

Vinti, J. P., New method of solution for unretarded satellite orbits ... 105

W

Walz, A., Compressible turbulent boundary layers with heat transfer and pressure gradient in flow direction ... 53
Water release, effect on the free outflow hydrograph ... 117
Water storage, analytical integration of the differential equation ... 43
Watson's lemma, modification ... 15
Weber parabolic cylinder functions of large orders, uniform asymptotic expansions ... 131

Y

Yerdjerich, V. M., Analytic integration of the differential equation for water storage ... 43
—, Effect of sudden water release on the reservoir free outflow hydrograph ... 117

Z

Zeros of certain polynomials ... 21

UNIVERSITY OF MICHIGAN #56-160074
PHYSICS LIBRARY
2035 RANDALL LAB.
ANN ARBOR MICH
D-18220

UNITED ST
GOVERNMENT PRINTING OFFICE
DIVISION OF PUBLIC DOCUMENTS
WASHINGTON 25, D.C.

OFFICIAL BUSINESS

PENALTY FOR PRIVATE USE TO AVOID PAYMENT OF POSTAGE, $300 (GPO)

National Bureau of Standards Journal of Research Vol. 63B, No. 2 October–December 1959
B. Mathematics and Mathematical Physics

Contents

Page

Applications of a theorem on partitioned matrices. E. V. Haynsworth. 73

Capacity requirement of a mail sorting device: II. A. J. Goldman. 79

Analytic comparison of suggested configurations for automatic mail sorting equipment. B. K. Bender and A. J. Goldman. 83

New method of solution for unretarded satellite orbits. John P. Vinti. 105

Effect of sudden water release on the reservoir free outflow hydrograph. Vujica M. Yevdjevich. 117

Uniform asymptotic expansions for Weber parabolic cylinder functions of large orders. F. W. J. Olver. 131

Publications of the National Bureau of Standards. 171

Index to volume 63B, July to December 1959. 175

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington 25, D.C. Price 75 cents (single copy). Subscription price: $2.25 a year; $0.50 additional for foreign mailing.

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1959